Estadística



Estrategias de Muestreo

Diseño de encuestas y estimación de parámetros

H. Andrés Gutiérrez Rojas

ediciones de la U

Estadística



Estrategias de Muestreo

Diseño de encuestas y estimación de parámetros

H. Andrés Gutiérrez Rojas

ediciones de la U

Bogotá - México, D.F.

Gutiérrez Rojas, H. Andrés

Estrategias de muestreo, diseño de encuestas y estimación de parámetros/ H. Andrés Gutiérrez Rojas, -- Bogotá : Ediciones de la U, 2016.
568 p. ; 24 cm.
ISBN 978-958-762-586-8 / e-ISBN 978-958-762-587-5
1. Estadística 2. Muestreo 3. Probabilidades I. Tít.
519.5 ed.

Área: Estadística
Primera edición: Bogotá, Colombia, agosto de 2016
ISBN. 978-958-762-586-8

Coordinación editorial: Adriana Gutiérrez M.
Carátula: Ediciones de la U
Impresión: Digiprint Editores SAS
Calle 63 bis #70-49, Pbx. (57+1) 4307050

Impreso y hecho en Colombia
Printed and made in Colombia



Apreciad@ cliente:

Es gratificante poner en sus manos estas obras, por esta razón le invitamos a que se registre en nuestra web: **www.edicionesdelau.com** y obtenga beneficios adicionales como:



- **Complementos digitales de las obras**
- **Actualizaciones de las publicaciones**
- **Interactuar con los autores a través del blog**
- **Información de nuevas publicaciones de su interés**
- **Noticias y eventos**

Adquiere nuestras publicaciones en formato e-book



Visítanos en:

www.edicionesdelau.com

Sus pedidos a:
Carrera 27 # 27-43 • Barrio Teusaquillo
PBX. (57-1) 4-065861 • (57-1) 4-942601 • Móvil: 310 - 6256033
comercial@edicionesdelau.com - gerencia@edicionesdelau.com
Bogotá - Colombia

Av. Coyoacán 1812 A. Acacias Benito Juárez C.P. 03240
PBX. (52) 55-63051703 • Cel. 044 5544439418
janethcr@gruporamadelau.com
México D. F. - México

A mi padre,
Hugo Gutiérrez Moreno,
en medio de la esperanza bienaventurada.

Prólogo

El muestreo es tal vez el oficio que caracteriza mejor la disciplina estadística. Se trata nada menos que de la planeación y ejecución de la recolección organizada de información con el fin de conocer o estimar propiedades generales de una población o de un fenómeno natural. Para ello, involucra, no solo la estrategia de selección probabilística de una muestra, sino que también hace explícito el estimador que proveerá el valor del parámetro que, a juicio del investigador, resume la propiedad estudiada.

Es indudable que los sistemas de cómputo han contribuido enormemente al desarrollo de nuevos métodos y a la práctica de los tradicionales con mayor confianza. Las fórmulas y los algoritmos se programan y se ponen en marcha y todo el trabajo numérico se deja en manos del computador, incluyendo la generación de los números pseudoaleatorios como punto de partida para la selección de las muestras y para la exploración de propiedades de los diseños mediante métodos de simulación. Igualmente, el muestreo se ha beneficiado de desarrollos abundantes y profundos de otros campos de la estadística que lo han enriquecido con nuevos métodos en los años recientes.

El profesor Andrés Gutiérrez emprende la tarea de entregar al lector un libro conceptualmente sólido, que se caracteriza por el equilibrio en la presentación intuitiva de los conceptos, apoyándose en (1) su estrategia pedagógica de Marco y Lucy, manteniendo una población y un marco de muestreo a lo largo de su texto, lo que le permite ejemplificar y comparar los diseños en lo referente a su eficiencia, (2) el desarrollo matemático riguroso de las propiedades de los diseños y de los estimadores, y (3) la práctica de los algoritmos con el paquete TeachingSampling, elaborado por él mismo para ilustrar los procedimientos propuestos.

De esta manera, el estudioso del tema encuentra cómo aplicar de manera inmediata los diversos procedimientos sin verse en la necesidad de emprender manualmente la búsqueda de soluciones o, si lo desea, tiene a su alcance los programas que le permiten comparar los resultados obtenidos con los que encuentra por sus propios medios. La primera parte contiene los métodos tradicionales que usualmente se enseñan en un primer curso de muestreo. La segunda está diseñada para un curso más avanzado de pregrado. En la tercera, el autor incluye temas recientes, provenientes de artículos publicados en revistas especializadas y de aparición más escasa en otros textos y que pueden ser parte de cursos de posgrado.

La lectura del texto es agradable y clara, acompañada de numerosos ejemplos lexicográficos con cantidades pequeñas de datos que ilustran el detalle de las posibilidades. Sin duda, el libro que nos presenta el profesor Andrés Gutiérrez tiene

el sello de un trabajo personal serio y novedoso que el lector apreciará.

Jorge Ortiz Pinilla, PhD.
Diplôme de docteur de troisième cycle
Université Henri Poincaré, Nancy 1

Prefacio

Aunque muy poderoso, el término estrategia de muestreo no ha tenido la trascendencia pertinente en el mundo del muestreo. Se habla de la eficiencia, precisión e incluso insesgamiento de un estimador sin tener en cuenta que tales propiedades están ligadas al diseño de muestreo que se haya utilizado en la recolección de la información. Para el autor, el aprendizaje de esta materia es más sencillo cuando se valora de igual manera el diseño de muestreo junto con el estimador del parámetro de interés utilizado en la población finita. No se puede desconocer la regla de oro del muestreo que clama: utilizar diseños de muestreo que induzcan probabilidades de inclusión (o selección, según sea el caso) proporcionales al valor de la característica de interés en la población y utilizar estimadores que involucren dichas probabilidades. Por lo anterior, se ha decido titular este texto como *Estrategias de muestreo, diseño de encuestas y estimación de parámetros.*

En la búsqueda combinada de una mejor estrategia de muestreo se ha querido dividir este texto en tres partes que pueden ser utilizadas en los distintos niveles de pregrado así como en cursos de posgrado, dependiendo de la dificultad del tema. La división del libro corresponde al desarrollo teórico del muestreo a través de su corta historia.

La primera parte del libro es un recuento de las estrategias de muestreo más comúnmente utilizadas. Con una muy robusta rigurosidad estadística y matemática, el lector será introducido en el ámbito de la inferencia basada en el diseño de muestreo, que considera a los valores de la característica de interés como pseudo-parámetros fijos y no como realizaciones de variables aleatorias. Este recorrido se hace más ameno al introducir a Marco y Lucy, compañeros inseparables en cada estrategia planteada, que corresponden a conjuntos de datos que se obtuvieron para la realización de una encuesta. Así, Marco es el pseudonombre del marco de muestreo y Lucy corresponde a una población de empresas en el sector industrial. A diferencia de la mayoría de los libros de muestreo clásicos, se ha querido proponer la revisión de todas las estrategias de muestreo con un sólo ≪marco≫ de muestreo, en algunas ocasiones más generoso que en otras, y una sola población, ≪Lucy≫, con el fin de plantear al lector un problema que puede ser resuelto desde diversos ángulos y no, como sucede en algunos libros de muestreo, proponer ejemplos resueltos en donde el lector no tiene la oportunidad de cuestionarse acerca del desarrollo de la estrategia. En esta parte se exponen tres conceptos totalmente determinantes al momento de plantear una estrategia de muestreo. El primero, y más importante de ellos, es el soporte que define la realización de una muestra probabilística y por consiguiente la validez en la inferencia. El lector puede darse cuenta de la importancia de este concepto al diferenciarlo de la muestra aleatoria,

que no es más que un vector aleatorio. En segundo lugar se hace énfasis en el concepto de diseño de muestreo tratándolo como una distribución de probabilidad multivariante sobre el soporte. En la mayoría de estrategias planteadas en esta primera parte se demuestra que efectivamente el diseño de muestreo propuesto verifica las propiedades de una distribución de probabilidad. Por supuesto, el último concepto es el de estimador cuya definición y uso es más difundido entre los usuarios. Cada capítulo y cada sección adjunta un pequeño ejemplo léxico-gráfico y una aplicación de la estrategia propuesta con Marco y Lucy mediante el desarrollo computacional del paquete TeachingSampling Gutierrez (2009) creado en el ambiente del software de uso libre R de la manera más amigable posible.

La segunda parte de este texto, corresponde a la utilización y aprovechamiento de la información auxiliar disponible en el marco de muestreo. No sólo se revisan estimadores que mejoran la eficiencia de la estrategia, sino que también se adjunta al proceso de estimación el uso de un modelo que permite *describir* el comportamiento de la característica de interés en la población dando un paso muy importante en el desarrollo de la inferencia al tratar, aunque de manera aislada, a la característica de interés como una variable aleatoria en el modelo propuesto. En esta parte, se sigue el hermoso enfoque de Bethelehem & Keller (1987) en donde no se hacen supuestos acerca de la validez del modelo de superpoblación. En esta línea de pensamiento la única aseveración que se realiza es que los residuales del modelo sí tienen una menor varianza que la característica de interés. Al final de esta parte se presenta una muy breve introducción a la inferencia en poblaciones finitas bajo un enfoque basado en el modelo supuesto. Es interesante observar que, con el pasar de los días, el uso de estas técnicas es más y más común, sobre todo en la estimación de dominios raros más conocidos como áreas pequeñas. Este apartado empieza con la famosa discusión de Basu (1971) que aun hoy sigue siendo la piedra de tropiezo entre las corrientes de la estadística. Este tipo de inferencia no considera el diseño de muestreo ni la forma en la que la información fue recolectada, sino que se basa en el modelo poblacional propuesto para realizar la inferencia. Sin embargo, cuando el modelo poblacional es errado también los serán las estimaciones.

La tercera parte, apta para un curso de posgrado, intenta acceder a los mayores avances metodológicos que, con el paso del tiempo, dejan de ser innovaciones para convertirse en técnicas obligatorias en la mejora de la eficiencia de la estrategia. Entre otros selectos temas, se consideran los estimadores de calibración, el muestreo balanceado y el muestreo indirecto. Estos contenidos están supeditados a un gran sesgo personal inducido por los años de asistencia al seminario de muestreo de la Universidad Nacional de Colombia.

Por supuesto, este libro no hubiera podido ser escrito sin la enorme influencia del maestro Leonardo Bautista, quien me enseñó que lo importante no es la memorización de fórmulas sino darles sentido y traerlas a la vida mediante el uso de la mejor estrategia. Lo anterior tiene sentido, no sólo en el desarrollo práctico de la ciencia estadística, sino en el diario vivir.

Por último, el autor expresa sus agradecimientos a Dios, quien le ha dado todo lo que tiene y le ha seleccionado para pertenecer en la muestra; de no ser así, no se habrían escrito estas líneas. Agradece el infinito apoyo que encontró en su esposa Hanwen Zhang en la corrección de las demostraciones así como en la invaluable motivación que le ha brindado desde el día que la conoció. De la misma manera,

agradece enormemente el continuo ánimo de su tío y amigo Oscar Ortega, por sus valiosos consejos desde su juventud y porque gran parte de este material se gestó en su residencia de los Estados Unidos. A su padre por la excelente crianza, a su abuela Lola porque sus oraciones nunca han cesado y a sus tías Martha, Myriam, Pilar y Clara. A su hermanita Tatiana Torres. A sus valiosos amigos.

Este texto es un producto del grupo de investigación en Muestreo y Marketing, adscrito al Centro de Investigaciones y Estudios Estadísticos (CIEES) de la Facultad de Estadística de la Universidad Santo Tomás.

Contenido

Parte I

Inferencia basada en el diseño de muestreo

Capítulo 1

Encuestas y estudios por muestreo

Durante todo el siglo pasado, ha surgido una serie de teorías y principios que ofrecen un marco de referencia unificado en el diseño, implementación y evaluación de encuestas. Este marco de referencia se conoce comúnmente como el paradigma del ≪error total de muestreo≫ y ha encaminado la investigación moderna hacia una mejor calidad de las encuestas.

Robert M. Groves (2004)

Este capítulo, a manera de introducción, busca identificar los principios (no matemáticos) del diseño, recolección, procesamiento y análisis de los estudios por muestreo, cuyo crecimiento va en aumento al pasar de los años, pero que sigue teniendo ciertas limitantes de tipo económico y logístico. Un estudio por muestreo involucrará a profesionales de diferentes disciplinas quienes se ocupan de la reducción de costos y el aumento de la calidad de las estimaciones. Un gran campo de la ciencia estadística se preocupa por minimizar los errores muestrales mientras que, por otra parte, otro gran campo de las ciencias sociales se ocupa en minimizar los errores que pueden ser cometidos en el periodo de la recolección de los datos.

1.1 Conceptos metodológicos

El muestreo es un procedimiento que responde a la necesidad de información estadística precisa sobre la población y los conjuntos de elementos que la conforman; el muestreo trata con investigaciones parciales sobre la población que apuntan a inferir a la población completa. Es así como en las últimas décadas ha tenido bastante desarrollo en diferentes campos principalmente en el sector gubernamental con la publicación de las estadísticas oficiales que permiten realizar un seguimiento a las metas del gobierno, en el sector académico, en el sector privado y de comunicaciones. Según Lohr (2000) el gasto anual en encuestas por muestreo en Estados Unidos representa de 2 a 5 billones de dólares. Este aumento del uso de las técnicas de muestreo en la investigación es claro porque es un procedimiento

que cuesta mucho menos dinero, consume menos tiempo y puede incluso ser más preciso que al realizar una enumeración completa, también llamada censo. Una muestra bien seleccionada de unos cuantos miles de individuos puede representar con gran precisión una población de millones.

Es requisito fundamental de una buena muestra que las características de interés que existen en la población se reflejen en la muestra de la manera más cercana posible, para esto se necesitan definir los siguientes conceptos

- **Población objetivo:** es la colección completa de todas las unidades que se quieren estudiar.
- **Muestra:** es un subconjunto de la población.
- **Unidad de muestreo:** es el objeto a ser seleccionado en la muestra que permitirá el acceso a la unidad de observación.
- **Unidad de observación:** es el objeto sobre el que finalmente se realiza la medición.
- **Variable de interés:** es la característica propia de los individuos sobre la que se realiza la inferencia para resolver los objetivos de la investigación.

En la teoría de muestreo la variable de interés no se supone como una variable aleatoria sino como una cantidad fija o una característica propia de las unidades que componen la población.

1.1.1 Encuesta

Por **encuesta** se entiende una investigación estadística con las siguientes características:

1. El objetivo de una encuesta es proveer información acerca de la población finita y/o acerca de subpoblaciones de interés especial.
2. Asociado con cada elemento de la población existe una o más variables de interés. Una encuesta permite conseguir información sobre características poblacionales desconocidas llamadas parámetros. Éstas son funciones de los valores de las variables de interés y son desconocidos y requeridos.
3. El acceso y observación de los elementos de la población se establece mediante un algoritmo de muestreo, que es un mecanismo que asocia los elementos de la población con unidades de muestreo.
4. Una muestra de elementos se escoge. Esto puede ser hecho mediante la selección de las unidades de observación en el esquema. Una muestra es probabilística si se realiza mediante un mecanismo probabilístico y se conoce la probabilidad de selección de todas las posibles muestras.
5. Los elementos seleccionados en la muestra son observados y se realiza el proceso de medición; es decir para cada elemento de la muestra la variable de interés se mide y sus valores se graban.

6. Los valores grabados de las variables son usados para calcular estimaciones de los parámetros de interés.

7. Las estimaciones son finalmente publicadas. Estas sirven para la toma de decisiones.

Ciclo de vida de una encuesta

Groves, Fowler, Couper, Lepkowski, Singer & R. (2004) afirman que una encuesta va desde el diseño, pasando por la ejecución hasta, la entrega de las estimaciones. Si no se realiza un buen diseño no habrán buenas estimaciones. En este camino, el investigador debe transitar los siguientes pasos:

1. **Búsqueda de constructores:** los constructores son las ideas abstractas acerca de las cuales el investigador desea inferir. En una encuesta de victimización, se busca medir cuántos incidentes relacionados con crímenes tuvieron lugar en cierto periodo de tiempo; el investigador debe decidir acerca de ¿qué es un crimen?, ¿quién es una víctima?. En una encuesta de calidad de vida, se desea saber cuántas personas pobres hay en una determinada región; por tanto, es necesario decidir acerca de ¿qué es pobreza?

2. **Medición:** la cuestión clave para realizar una buena medición es diseñar preguntas que produzcan respuestas que reflejen perfectamente los constructores que se intentan medir. Por ejemplo, en la encuesta de victimización, se puede preguntar lo siguiente: ≪en los últimos seis meses ¿ha llamado usted a la policía para reportar algo que le haya sucedido y que usted considere que sea un crimen?≫. Por otro lado, en la encuesta de calidad de vida, un indicador de pobreza puede estar dado en términos del número de electrodomésticos que posee el hogar. Así, es posible preguntar lo siguiente: ≪¿cuántos televisores tiene en su hogar?≫ o también ≪¿cuántas bombillas eléctricas tiene su hogar?≫

3. **Respuesta:** la naturaleza de las respuestas está determinada por la naturaleza de las preguntas. En algunas ocasiones la respuesta puede ser parte de la pregunta, siendo la tarea del respondiente escoger entre las categorías preguntadas; en otras ocasiones, el respondiente genera una respuesta concreta en sus propias palabras.

4. **Edición:** existen relaciones lógicas entre las preguntas de una encuesta. Por ejemplo, si el respondiente declara tener 12 años de edad y haber dado a luz a 5 hijos, debe existir un proceso de edición para este individuo. Este proceso intenta detectar datos atípicos y revisar la información para obtener la mejor medida del constructor buscado.

5. **Análisis y entrega de resultados:** el proceso estadístico arroja estimaciones que permiten la toma de decisiones y la resolución de los objetivos propuestos al comienzo de la investigación.

1.1.2 Marco de muestreo

Todo procedimiento de muestreo probabilístico requiere de un dispositivo que permita identificar, seleccionar y ubicar a todos y cada uno de los objetos pertenecientes a la población objetivo y que participarán en la selección aleatoria. Este dispositivo se conoce con el nombre de **marco de muestreo**. En investigaciones por muestreo se consideran dos tipos de objetos:

- **Elementos:** las unidades básicas e individuales sobre las que se realiza la medición.
- **Conglomerado:** agrupación de elementos cuya característica principal es que son homogéneos dentro de sí, y heterogéneos entre sí.

Cuando se dispone de un marco de elementos, se puede aplicar un diseño de muestreo de elementos; en muchas ocasiones se utilizan diseños de muestreo de conglomerados aunque se disponga de un marco de elementos. Si no se dispone de un marco de elementos (o es muy costoso construirlo) se debe recurrir a diseños de muestreo en conglomerados; es decir, que se utilizan marcos de conglomerados. Por ejemplo, al realizar una encuesta cuya unidad de observación sean las personas que viven en una ciudad, es muy difícil poder acceder a un marco de muestreo de las personas. Sin embargo, se puede tener acceso a la división sociodemográfica de la ciudad y así seleccionar algunos barrios de la ciudad, en una primera instancia y luego, seleccionar a las personas de los barrios en una segunda instancia. En el ejemplo anterior, los barrios son un ejemplo claro de conglomerados. Estas agrupaciones de elementos tienen la características de aparecer en el estado de la naturaleza. De esta forma, si se dispone de un marco de elementos, por ejemplo, el listado de empleados de una entidad, es posible aplicar un diseño de muestreo de elementos, realizar la selección aleatoria y de acuerdo a ese mismo diseño realizar las estimaciones necesarias. El lector debe recordar que los elementos son las entidades que componen la población y las unidades de muestreo son las entidades que conforman el marco muestral. Cuando no existe un marco de muestreo disponible es necesario construirlo. Existen dos tipos de marcos de muestreo, a saber:

- **De Lista:** listados físicos o magnéticos, ficheros, archivos de expedientes, historias clínicas que permiten identificar y ubicar a los objetos que participarán en el sorteo aleatorio.
- **De Área:** mapas de ciudades y regiones en formato físico o magnético, fotografías aéreas, imágenes de satélite o similares que permiten delimitar regiones o unidades geográficas en forma tal que su identificación y su ubicación sobre el terreno sea posible.

Es una virtud del marco si contiene **información auxiliar** que permite aplicar diseños muestrales y/o estimadores que conduzcan a estrategias más eficientes con respecto a la precisión de los resultados. O también si la información auxiliar[1]

[1] Toda información auxiliar disponible para todos y cada uno de los elementos del universo afecta directamente la estrategia empleada para obtener los objetivos de la investigación. Con respecto a la información auxiliar, es deseable que esté bien correlacionada con la característica de interés.

está organizada por órdenes deseables. Se llama información auxiliar **discreta**, si el marco de muestreo permite la desagregación de la población objetivo en categorías o grupos poblacionales más pequeños. Por ejemplo nivel socioeconómico, grupo industrial, etc. Se llama información auxiliar **continua** si existe una o varias características de interés de tipo continuo y positivas. Es deseable que la información auxiliar continua esté altamente relacionada con la característica de interés.

Por otra parte, un marco de muestreo es defectuoso si presenta alguno o varios de los siguientes casos:

- **Sobre-cobertura:** se presenta si en el dispositivo aparecen objetos que no pertenecen a la población objetivo. *No son todos los que están.*
- **Sub-cobertura:** se da cuando algunos elementos de la población objetivo no aparecen en el marco de muestreo o cuando no se ha actualizado la entrada de nuevos integrantes. *No están todos los que son.*
- **Duplicación:** La duplicación en un marco de muestreo se presenta si en el dispositivo aparecen varios registros para un mismo objeto. La razón más frecuente para la presencia de este defecto es la construcción no cuidadosa del marco a partir de la unión de registros administrativos de dos o más fuentes de información.

Estos defectos ocasionan errores en el cálculo de las expresiones que se utilizarán para generar las correspondientes estimaciones, generando sesgo, pérdida de precisión y, en algunos casos, que los resultados del estudio pierdan toda validez.

Tipos de poblaciones objetivo

Groves, Fowler, Couper, Lepkowski, Singer & R. (2004) consideran que los tipos de poblaciones objetivo que se presentan de manera más frecuente en un estudio por muestreo son las siguientes

- **Hogares y personas:** el marco de muestreo más utilizado en estas poblaciones es de área. Como está basada en zonas geográficas, este tipo de marco requiere la vinculación de los hogares o personas a cada una de las áreas. Cuando se requiere seleccionar personas, este tipo de marcos hace necesarias muchas etapas de muestreo; de esta forma, se selecciona un subconjunto de zonas geográficas. Para cada zona seleccionada, se procede a seleccionar un subconjunto de secciones, luego de manzanas, luego de hogares y, finalmente, para cada hogar se seleccionan las personas; siendo éstas las unidades de observación.
- **Clientes, empleados o miembros de organizaciones:** por lo general, para la selección de miembros de organizaciones se manejan marcos de lista. Es importante que el estadístico esté al tanto de la frecuencia y manera de actualización de la lista pues pueden presentar los tres tipos de defectos vistos anteriormente.
- **Organizaciones:** existen diversos tipos de organizaciones, como por ejemplo, iglesias, prisiones, empresas, hospitales, escuelas, etc. En encuestas a

establecimientos comerciales, es frecuente tener acceso a marcos de lista que agrupan a negocios con gran dispersión entre sí. Así, se puede encontrar desde la tienda de barrio, cuyas ventas ascienden a 1000 dólares al mes, hasta un hipermercado que vende 500 millones de dólares al mes.

- **Eventos:** en algunas ocasiones, la población objetivo son eventos. Hay muchos tipos de eventos que clasifican para la realización de una encuesta; entre ellos están los matrimonios, nacimientos, fallecimientos, periodos de depresión, tránsito de un automóvil en un segmento de la vía. Los marcos de muestreo para los eventos, de manera frecuente, son marcos de personas. Así, una persona ya ha experimentado el evento o no. De hecho, puede haber experimentado varios eventos. Sin embargo, otro marco de muestreo para eventos puede estar dado en periodos de tiempo o espacio.

- **Poblaciones poco frecuentes:** cuando la incidencia es muy baja (por ejemplo las poblaciones de invidentes o con alguna enfermedad rara). Generalmente, la manera para acceder a este tipo de poblaciones es mediante un marco de muestreo que contenga a esta población como un subconjunto de elementos que pueden ser ubicados.

Ejemplo 1.1.1. Suponga que una entidad oficial del gobierno de su país está interesada en la realización de una encuesta de desempleo con el fin de determinar a) cuántas personas actualmente pertenecen a la fuerza laboral, tanto en el país en cuestión como en sus regiones o subdivisiones geográficas y b) qué proporción de éstas están desempleadas. Con base en lo anterior se tienen los siguientes aspectos para la realización de dicho estudio:

- *Población objetivo:* Todas las personas de Colombia.

- *Dominios o subgrupos de interés:* Grupos de edad, género, grupos ocupacionales y regiones del país.

- *Características de interés:* Pertenencia a la fuerza laboral y estado de empleo. Éstas toman valor uno o cero.

- *Parámetros de interés:* Número total de personas pertenecientes a la fuerza laboral, número total de desempleados, proporción de desempleo.

- *Muestra:* Se selecciona una muestra de la población con la ayuda de mecanismos de identificación y ubicación de las personas en el país.

- *Observaciones:* Cada persona incluida en la muestra es visitada por un encuestador entrenado, quien hará preguntas siguiendo un cuestionario estandarizado y recolectará las respuestas en un instrumento apropiado.

- *Procesamiento:* Los datos se editan y se preparan para la etapa de estimación.

- *Estimación:* Se calculan las estimaciones sobre los parámetros de interés y también indicadores acerca de la incertidumbre de estas estimaciones.

1.1.3 Sesgo

En el diseño y puesta en marcha de una encuesta puede ocurrir cierto tipo de situaciones que pueden sesgar las estimaciones finales. Este tipo de sesgos puede ocurrir antes, durante y después de la recolección de los datos. Es tarea del estadístico advertir ante todas las posibles instancias de los problemas que causan los sesgos y procurar que, en todas las etapas de la encuesta, se minimice el error humano y el error estadístico para que al final los resultados del estudio sean tan confiables como sea posible.

Sesgo de selección

Este tipo de sesgo ocurre cuando parte de la población objetivo no está en el marco de muestreo. Una muestra a conveniencia[2] es sesgada pues las unidades más fáciles de elegir o las que más probablemente respondan a la encuesta no son representativas de las unidades más difíciles de elegir. (Lohr 2000) afirma que se presenta este tipo de sesgo si:

1. La selección de la muestra depende de cierta característica asociada a las propiedades de interés. Por ejemplo: Frecuencia con que los adolescentes hablan con los padres acerca del SIDA.

2. La muestra se realiza mediante elección deliberada o mediante un juicio subjetivo. Por ejemplo, si el parámetro de interés es la cantidad promedio de gastos en compras en un centro comercial y el encuestador elige a las personas que salen con muchos paquetes, entonces la información estaría sesgada puesto que no está reflejando el comportamiento promedio de las compras.

3. Existen errores en la especificación de la población objetivo. Por ejemplo, en encuestas electorales, cuando la población objetivo contiene a personas que no están registradas como votantes ante la organización electoral de su país.

4. Existe sustitución deliberada de unidades no disponibles en la muestra. Si, por alguna razón, no fue posible obtener la medición y consecuente observación de la característica de interés para algún individuo en la población, la sustitución de este elemento debe hacerse bajo estrictos procedimientos estadísticos y no debe ser subjetiva en ningún modo.

5. Existe ausencia de respuesta. Este fenómeno puede causar distorsión de los resultados cuando los que no responden a la encuesta difieren críticamente de los que si respondieron.

6. La muestra está compuesta por respondientes voluntarios. Los foros radiales, las encuestas de televisión y los estudios de portales de internet no proporcionan información confiable.

[2]A pesar de que las muestras por conveniencia o por juicio no pueden ser utilizadas para estimar parámetros de la población, éstas sí pueden proporcionar información valiosa en las primeras etapas de una investigación o cuando no es necesario generalizar los resultados a la población.

Sesgo de medición

Este tipo de sesgo ocurre cuando el instrumento con el que se realiza la medición tiene una tendencia a diferir del valor verdadero que se desea averiguar. Éste sesgo debe ser considerado y minimizado en la etapa de diseñó de la encuesta. Nótese que ningún análisis estadístico puede revelar que una pesa añadió a cada persona 2Kg de más en un estudio de salud. (Lohr 2000) cita algunas situaciones en donde se presenta este sesgo de medición:

1. Cuando el respondiente miente. Esta situación se presenta a menudo en encuestas que pregunta acerca del ingreso salarial, alcoholismo y drogadicción, nivel socioeconómico e incluso edad.
2. Difícil comprensión de las preguntas. Por ejemplo: ¿No cree que no este es un buen momento para invertir? La doble negación en la pregunta es muy confusa para el respondiente.
3. Las personas tienden a olvidar. Es bien sabido que las malas experiencias suelen ser olvidadas; esta situación debe acotarse si se está trabajando en una encuesta de criminalidad.
4. Distintas respuestas a distintos entrevistadores. En algunas regiones es muy probable que la raza, edad o género del encuestador afecte directamente la respuesta del entrevistado.
5. Leer mal las preguntas o polemizar con el respondiente. El encuestador puede influir notablemente en las respuestas. Por lo anterior, es muy importante que el proceso de entrenamiento del entrevistador sea riguroso y completo.
6. La muestra está compuesta por respondientes voluntarios. Los foros radiales, las encuestas de televisión y los estudios de portales de internet no proporcionan información confiable.

1.2 Marco y Lucy

Este libro toma como base de aplicación una investigación gubernamental que quiere responder al objetivo de *medir el crecimiento económico en el sector industrial.*

Suponga que para completar el objetivo se ha propuesto desarrollar una encuesta a las empresas que hacen parte del sector industrial, para conocer el comportamiento del sector en términos de **constructores** financieros, sociales y fiscales. Una vez termine el proceso de medición, se pueden calcular estimaciones y construir indicadores que permitan inferir acerca del crecimiento del sector en el periodo de interés.

La **población objetivo** la conforman todas las empresas cuya actividad principal esté ligada al sector industrial. El proceso de medición se hará con base en las **características de interés**; a saber: ingresos en el último año fiscal, impuestos declarados en el último año fiscal y número de empleados. Adicionalmente, se requiere conocer si la empresa envía periódicamente algún tipo de material

publicitario por correo electrónico porque se sospecha que las empresas obtienen más ingresos cuando utilizan esta estrategia publicitaria, lo cual es favorable para el gobierno porque aumenta la contribución impositiva y aumenta la creación de empleos.

Para obtener las respuestas, un entrevistador visitará las instalaciones físicas de la empresa y realizará las siguientes preguntas:

1. En el último año fiscal, ¿a cuánto ascendieron los ingresos en esta empresa?
2. En el último año fiscal, ¿a cuánto ascendieron los impuestos declarados por esta empresa?
3. Actualmente, ¿cuántos empleados laboran para esta empresa?
4. ¿Esta empresa acostumbra a enviar periódicamente material publicitario por correo electrónico a sus clientes o potenciales clientes?

Se sabe que el tamaño de la población es de 2396 empresas. Dependiendo de la estrategia de muestreo que se vaya a utilizar y de la calidad del marco de muestreo, las unidades de muestreo pueden ser las mismas empresas.

Para abordar la selección de una muestra que permita la inferencia acerca del crecimiento económico del sector, se dispone de un marco de muestreo con las siguientes características para cada empresa que conforma la población.

1. **Identificador:** es una secuencia alfanumérica de dos letras y tres dígitos. Este número de identificación se le otorga a cada empresa en el momento de la constitución legal ante la entidad de registro pertinente.
2. **Ubicación:** es la dirección que se encuentra registrada en la declaración de impuestos.
3. **Zona:** la ciudad está conformada por barrios o zonas geográficas. Dependiendo de la dirección, la empresa pertenece a una y sólo una zona geográfica de la ciudad.
4. **Nivel:** según los registros tributarios, las empresas se catalogan en tres grupos:
   (a) Grandes: empresas que tributan 49 millones de dólares al año o más.
   (b) Medianas: empresas que tributan más de 11 millones y menos de 49 millones de dólares al año.
   (c) Pequeñas: empresas que tributan 11 millones de dólares al año o menos.

   Nótese que una empresa sólo puede pertenecer a un sólo un nivel industrial.

Visualización en R

El paquete `TeachingSampling` de R incluye dos archivos de datos. El marco de muestreo llamado `Marco` del cual se extraerá una muestra aleatoria de empresas que deben ser entrevistadas y que contiene la identificación, ubicación, zona y nivel de

cada una de las empresas del sector industrial. Por otro lado, incorpora el conjunto de datos llamado `Lucy` en donde, se encuentran los valores de las características de interés para todos los elementos de la población.

Para tener acceso a los dos conjuntos de datos es necesario cargar el paquete en el entorno de R. El paquete `TeachingSampling` puede ser cargado fácilmente mediante el uso de la siguiente instrucción:

```
> library(TeachingSampling)
```

Una vez cargado el paquete `TeachingSampling`, la visualización del marco de muestreo, se realiza de la siguiente forma:

```
> data(Marco)
> Marco[1:10,]

      ID Ubication Level Zone
1  AB001      c1k1 Small    A
2  AB002      c1k2 Small    A
3  AB003      c1k3 Small    A
4  AB004      c1k4 Small    A
5  AB005      c1k5 Small    A
6  AB006      c1k6 Small    A
7  AB007      c1k7 Small    A
8  AB008      c1k8 Small    A
9  AB009      c1k9 Small    A
10 AB010     c1k10 Small    A
```

La instrucción `Marco[1:10,]` se utiliza para mostrar las diez primeras empresas del marco de muestreo. Si se quiere visualizar todo el conjunto de datos, la instrucción `Marco` mostrará la totalidad del marco de muestreo. La función `names` muestra cada uno de los objetos que componen el archivo de datos, mientras que la función `dim` muestra las dimensiones del conjunto de datos.

```
> names(Marco)
[1] "ID"        "Ubication" "Level"     "Zone"
> dim(Marco)
[1] 2396    4
```

El tamaño poblacional es igual a 2396, por tanto el marco de muestreo tendrá igual número de filas y 4 columnas conteniendo la información acerca de la identificación y ubicación de cada empresa en la población. La función `attach` permite el acceso a cada una de las variables del conjunto de datos. Así, para tener acceso al identificador y/o ubicación de las diez primeras empresas del marco de muestreo basta con digitar el siguiente código

```
> attach(Marco)
> ID[1:10]
 [1] AB001 AB002 AB003 AB004 AB005 AB006 AB007 AB008 AB009 AB010
> Ubication[1:10]
 [1] c1k1  c1k2  c1k3  c1k4  c1k5  c1k6  c1k7  c1k8  c1k9  c1k10
```

De igual forma, se tiene acceso al valor de las características de interés para las diez primeras empresas del marco de muestreo, así

```
> data(Lucy)
> Lucy[1:10,]

      ID Ubication Level Zone Income Employees Taxes SPAM
1  AB001      c1k1 Small    A    281        41   3.0   no
2  AB002      c1k2 Small    A    329        19   4.0  yes
3  AB003      c1k3 Small    A    405        68   7.0   no
4  AB004      c1k4 Small    A    360        89   5.0   no
5  AB005      c1k5 Small    A    391        91   7.0  yes
6  AB006      c1k6 Small    A    296        89   3.0   no
7  AB007      c1k7 Small    A    490        22  10.5  yes
8  AB008      c1k8 Small    A    473        57  10.0  yes
9  AB009      c1k9 Small    A    350        84   5.0  yes
10 AB010     c1k10 Small    A    361        25   5.0   no
```

La lectura del archivo de datos se hace de la siguiente manera: tomando como referencia la fila número 3 (la tercera empresa del conjunto de datos), es una empresa cuyo número de identificación es `AB003`, ubicada en la dirección `c1k3`, de nivel industrial `Pequeño`, localizada en la zona `A`. Esta empresa registró en el último año fiscal un ingreso neto de 405 millones de dólares y realizó un tributo de 7 millones de dólares, actualmente da empleo a 68 empleados y sí envía periódicamente publicidad a sus clientes o potenciales clientes mediante correo electrónico.

Nótese que el conjunto de datos poblacionales `Lucy` contiene el valor de las características de interés para cada empresa. Hasta este momento no se ha seleccionado ninguna muestra, pero si se supone hipotéticamente que la muestra seleccionada hubiese sido las diez primeras empresas del marco de muestreo, la base de datos, después de la medición se vería como lo muestra la salida anterior y con estos datos se procede a realizar las estimaciones requeridas para el cumplimiento de los objetivos de la investigación.

Las estadísticas concernientes a las variables en las población se visualizan fácilmente con la función `summary` aplicada al conjunto de datos `Lucy`.

```
> summary(Lucy)
    Income          Employees          Taxes           SPAM
Min.   :   1.0   Min.   :  1.00   Min.   :  0.50   no : 937
1st Qu.: 230.0   1st Qu.: 38.00   1st Qu.:  2.00   yes:1459
Median : 390.0   Median : 63.00   Median :  7.00
Mean   : 432.1   Mean   : 63.42   Mean   : 11.96
3rd Qu.: 576.0   3rd Qu.: 84.00   3rd Qu.: 15.00
Max.   :2510.0   Max.   :263.00   Max.   :305.00
```

Por medio de la función `total`, tenemos acceso al total de las tres características de interés.

```
> total <- function(x){length(x)*mean(x)}
```

```
> attach(Lucy)
> total(Income); total(Employees); total(Taxes)

  Income      Employees   Taxes
  1035217     151950      28653.5
```

El sector industrial tiene altos ingresos que ascienden a 1.035.217 millones de dólares, aporta al gobierno 28.653 millones de dólares en tarifas impositivas y emplea un total de 151.950 personas.

La función `tapply` permite aplicar la función `total` y la función `mean` para calcular el total y el promedio, respectivamente, de las variables de interés en cada categoría de la variable `Level`. La función `table` hace un recuento del total de casos para una o más variables categóricas.

```
> tapply(Income,Level,total)
   Big   Medium    Small
103706   487351   444160
> table(SPAM,Level)
     Level
SPAM    Big     Medium  Small
  no    26      291     620
  yes   57      446     956
```

Nótese que la mayoría del ingreso del sector industrial es adquirido por las empresas medianas y pequeñas. Sin embargo, en promedio las empresas grandes doblan el ingreso de las medianas que a su vez es tres veces el ingreso de las empresas pequeñas. En términos absolutos, la estrategia publicitaria de enviar SPAM a los clientes o potenciales clientes se implementa con mayor frecuencia en las empresas pequeñas.

La función `xtabs` permite realizar una tabulación cruzada entre las variables categóricas `Level` y `SPAM` de la base de datos. Los datos de las celdas indican el total de la variable `Income`. Nótese que el ingreso de las empresas que utilizan el SPAM como estrategia de publicidad dobla el ingreso de las empresas que no utilizan SPAM en casi todos los niveles industriales.

```
> xtabs(Income~Level+SPAM)
        SPAM
Level          no        yes
  Big       31914      71792
  Medium   190852     296499
  Small    175186     268974
```

La figura 1.1. muestra la dispersión y locación de las características de interés por cada nivel industrial. En general, las empresas grandes tienen ingresos más altos, aportan una carga impositiva más alta y emplean a más personas que las empresas medianas y pequeñas. Es deseable que el marco de muestreo contenga la pertenencia al nivel industrial de cada empresa en la población porque es un buen discriminante y permite la implementación de estrategias de muestreo adecuadas

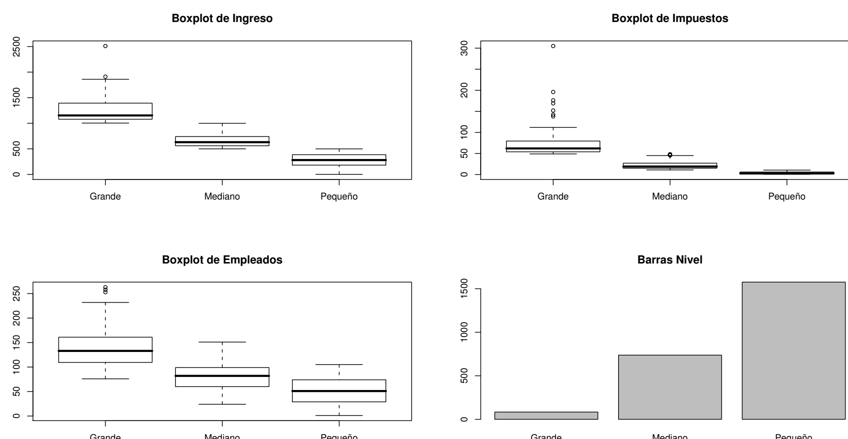


Figura 1.1: *Boxplot de las características de interés en cada nivel industrial.*

que guíen a estimaciones más precisas. La función `boxplot` permite la creación del diagrama de caja y la función `barplot` muestra un diagrama de barras del total de la variable `Level`.

```
> boxplot(Income  ~ Level,main=c("Boxplot de Ingreso"))
> barplot(table(Level),main=c("Barras Nivel"))
```

La figura 1.3. muestra que la distribución de las características de interés no es simétrica y es sesgada a la izquierda. Estos rasgos particulares se deben tener en cuenta al momento de escoger la mejor estrategia de muestreo. La función `hist` permite la creación de los histogramas y la función `pie` permite la creación de un gráfico de torta.

```
> hist(Income,main=c("Histograma de Ingreso"))
> pie(table(Level),main=c("Pie de Nivel"))
```

La correlación lineal entre las características de interés es alta; entre `Income` y `Taxes` existe una correlación de 0.91, esto se puede explicar porque las empresas tributan una mayor cantidad de dinero si han obtenido mayores ingresos y viceversa. Se utiliza la función `cor` para obtener la matriz de correlación entre las características de interés.

```
> Datos <- data.frame(Income, Employees, Taxes)
> cor(Datos)
            Income Employees    Taxes
Income    1.000000  0.645536 0.916954
Employees 0.645536  1.000000 0.646855
Taxes     0.916954  0.646855 1.000000
```

Para visualizar la relación entre las variables de interés, se utiliza la función `pairs` para obtener los diagramas de dispersión para cada par de variables justo como lo muestra la figura 1.4.

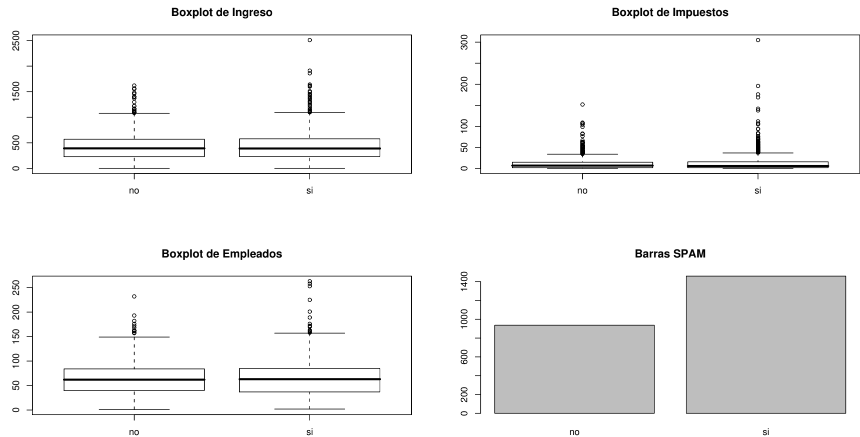


Figura 1.2: *Boxplot de las características de interés para cada comportamiento publicitario.*

La tabla 1.1. resume los parámetros de interés que, mediante una adecuada estrategia de muestreo, se deben estimar para resolver el objetivo principal de la investigación. Si se desean estimaciones discriminadas por nivel industrial, entonces la tabla 1.2. da cuenta del valor de estos parámetros dentro de los subgrupos poblacionales.

Consecuentemente, si se quieren estimaciones discriminadas por comportamiento publicitario, entonces la tabla 1.3. muestra el valor de cada uno de estos parámetros. Por último, si se buscan estimaciones discriminadas tanto por comportamiento publicitario cruzado con nivel industrial, entonces se cuenta con la tabla 1.4. que resume dicha información.

Tabla 1.1: *Parámetros de la población.*

	Ingreso	Impuestos	Empleados
N total	2.396	2.396	2.396
Suma	1.035.217	28.654	151.950
Media	432	12	63

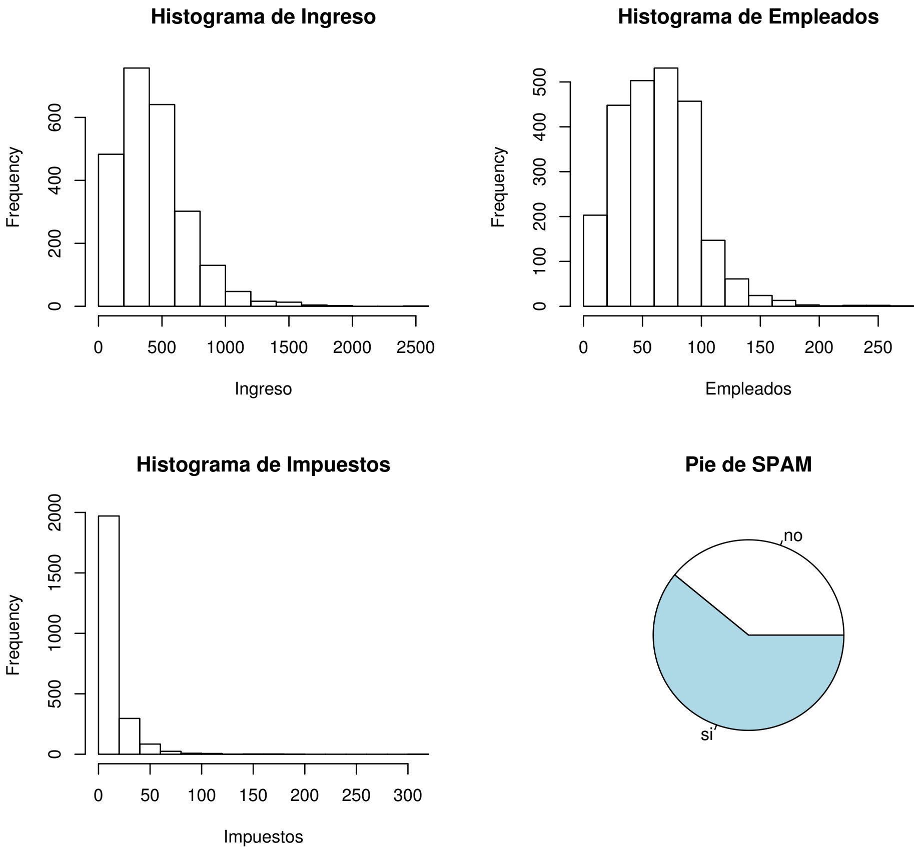


Figura 1.3: *Histograma de las características de interés.*

Tabla 1.2: *Parámetros de la población discriminados por nivel industrial.*

			Ingreso	Impuestos	Empleados
Nivel	Grande	N total	83	83	83
		Suma	103.706	6.251	11.461
		Media	1.249	75	138
	Mediano	N total	737	737	737
		Suma	487.351	16.293	59.643
		Media	661	22	81
	Pequeño	N total	1.576	1.576	1.576
		Suma	444.160	6.110	80.846
		Media	282	4	51

Tabla 1.3: *Parámetros de la población discriminados por comportamiento publicitario.*

			Ingreso	Impuestos	Empleados
SPAM	no	N total	937	937	937
		Suma	397.952	10.593	59.600
		Media	425	11	64
	si	N total	1.459	1.459	1.459
		Suma	637.265	18.061	92.350
		Media	437	12	63

Tabla 1.4: *Parámetros de la población discriminados por nivel industrial y por comportamiento publicitario.*

		SPAM					
		no			si		
		N total	Suma	Media	N total	Suma	Media
Grande	Ingreso	26	31.914	1.227	57	71.792	1.260
	Impuestos	26	1.844	71	57	4.4.07	77
	Empleados	26	3.587	138	57	7.874	138
Mediano	Ingreso	291	190.852	656	446	296.499	665
	Impuestos	291	6.322	22	446	9.971	22
	Empleados	291	23.745	82	446	35.898	80
Pequeño	Ingreso	620	175.186	283	956	268.974	281
	Impuestos	620	2.427	4	956	3.683	4
	Empleados	620	32.268	52	956	48.578	51

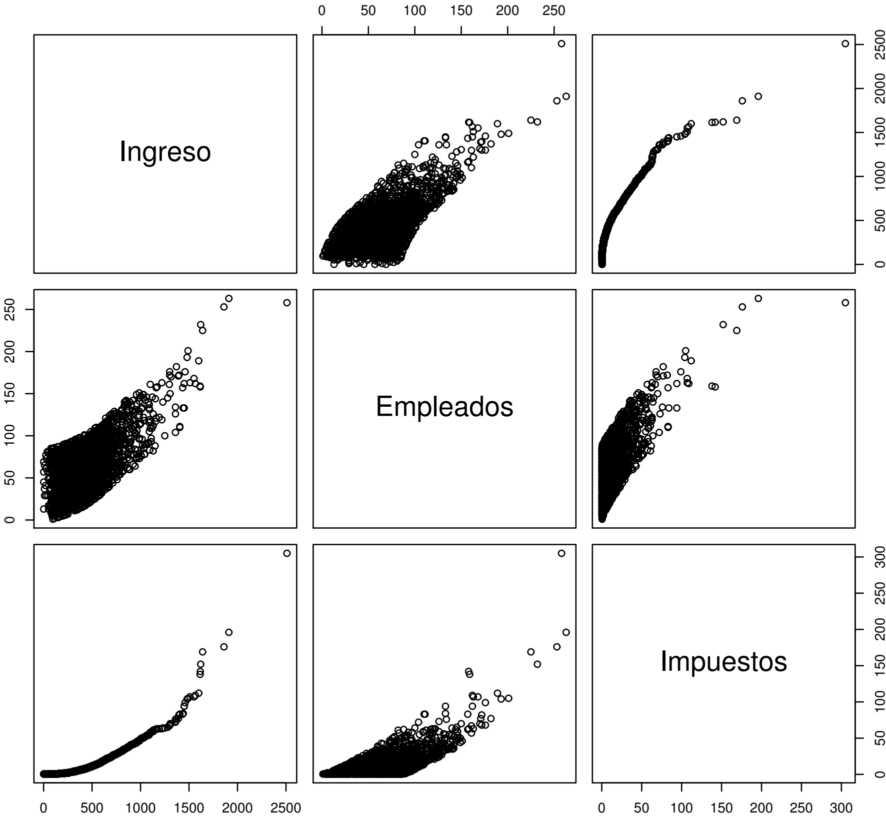


Figura 1.4: *Relación entre las características de interés.*

Capítulo 2

Muestras probabilísticas y estimadores

> La base matemática para el desarrollo del modelo de muestreo se encuentra en la teoría de la inferencia estadística y de manera más directa en la aplicación de los principios básicos de la teoría de probabilidad. Los resultados del modelo de muestreo sólo son válidos si se parte de la certeza de contar con una muestra que satisfaga las condiciones exigidas por la inferencia estadística.
>
> Leonardo Bautista (1998)

2.1 Población y muestra aleatoria

El proceso de estimación e inferencia en poblaciones finitas, que finalmente son las que fácilmente encontramos en la realidad y en las que se enfoca el muestreo, es muy diferente al proceso de inferencia de la estadística clásica. Esta última se trata a los valores observados como realizaciones de una variable aleatoria. En contravía con lo anterior, el muestreo asume que los valores observados corresponden a parámetros fijos poblacionales. Partiendo de este hecho formalicemos algunos conceptos que son de vital importancia en el estudio y análisis del muestreo.

2.1.1 Población finita

Definición 2.1.1. *Una* ***población finita*** *es un conjunto de N elementos $\{e_1, e_2, ..., e_N\}$. Cada unidad puede ser identificada sin ambigüedad por un conjunto de rótulos. Sea $U = \{1, 2, ..., N\}$ el conjunto de rótulos de la población finita. El tamaño de la población no es necesariamente conocido.*

Es el conjunto de N, donde $N < \infty$, unidades que conforman el universo de estudio. N es comúnmente llamado el tamaño poblacional. Cada elemento perteneciente a la población puede ser identificado por un rótulo. Sea U el conjunto de rótulos, tal que

$$U = \{1, ..., k, ..., N\}.$$

Se utilizará el subíndice k para denotar la existencia física del k-ésimo elemento. Nótese que el **tamaño de la población**, N, no siempre es conocido y en algunas ocasiones el objetivo de la investigación es poder estimarlo.

2.1.2 Muestra aleatoria

Es un subconjunto de la población que ha sido extraído mediante un mecanismo estadístico de selección. Notaremos con una letra mayúscula S a la muestra aleatoria[1] y con una letra minúscula s a una realización de la misma. De tal forma que, sin ambigüedad, una muestra seleccionada (realizada) es el conjunto de unidades pertenecientes a

$$s = \{1, ..., k, ..., n(S)\}.$$

El número de componentes de s es llamado el **tamaño de muestra** y no siempre es fijo. Es decir, en algunos casos $n(S)$ es una cantidad aleatoria. El conjunto de todas las posibles muestras se conoce como **soporte**. Haciendo una analogía con la inferencia estadística clásica, el soporte generado por una muestra aleatoria corresponde al espacio muestral generado por una variable aleatoria.

La anterior definición de muestra, en donde los elementos incluidos se listan dentro de un conjunto, corresponde a la forma clásica de notación. Sin embargo, una muestra también puede ser notada como un vector de tamaño N. De esta manera, la k-ésima entrada del vector denotará el número de veces que el elemento fue incluido o seleccionado; si el valor es cero, indica que el elemento no fue incluido en la muestra seleccionada; si el valor es distinto de cero, indica que el elemento sí fue seleccionado. Aunque ambas formas de notación tienen la misma interpretación, para evitar confusiones, se denotará la muestra en forma de vector con una $\mathbf{s}$ en negrilla, mientras que la muestra en forma de conjunto se denotara con una s simple sin negrilla. A continuación se dan definiciones más precisas acerca de la muestra aleatoria con o sin reemplazo.

Muestra aleatoria sin reemplazo

Definición 2.1.2. *Una* ***muestra sin reemplazo*** *se denota mediante un vector columna*

$$\mathbf{s} = (I_1, I_2, ..., I_N)' \in \{0, 1\}^N \tag{2.1.1}$$

donde

$$I_k = \begin{cases} 1 & \text{si el } k\text{-ésimo elemento pertenece a la muestra,} \\ 0 & \text{en otro caso} \end{cases} \tag{2.1.2}$$

Una muestra aleatoria se dice sin reemplazo si la inclusión de cada uno de los elementos se hace entre los elementos que no han sido escogidos aún; de esta

[1] Nótese que S es una variable aleatoria.

manera el conjunto s nunca tendrá elementos repetidos. El tamaño de muestra corresponde a la cardinalidad de s.

$$n(S) = \sum_{k \in U} I_k. \quad (2.1.3)$$

Como $n(S)$ no es una cantidad fija, es posible que ocurran uno de los siguientes escenarios: a) que la muestra no contenga a ningún elemento, entonces esta muestra se dice vacía; b) que la muestra contenga a todos los elementos de la población, esta muestra se conoce con el nombre de **censo**.

Muestra aleatoria con reemplazo

Definición 2.1.3. *Una **muestra con reemplazo** se denota mediante un vector columna*

$$\mathbf{s} = (n_1, n_2, ..., n_N)' \in \mathbb{N}^N \quad (2.1.4)$$

donde n_k es el número de veces que el elemento k está en la muestra

En algunos casos, por conveniencia del mecanismo de selección, el usuario prefiere tomar una muestra aleatoria con reemplazo si la inclusión de cada uno de los elementos tiene en cuenta a todos los elementos, ya sea que hayan sido escogidos para pertenecer en la muestra o no. De esta forma, el usuario puede seleccionar una muestra cuyo proceso de selección incluya a un individuo m veces (nótese que m puede ser mayor que N). Sin embargo, en una muestra aleatoria con reemplazo, dos o más componentes pueden ser idénticos. Un elemento que esté incluido más de una vez en s es llamado **elemento repetido**.

En principio el tamaño de muestra está dado por

$$n(S) = m = \sum_{k \in U} n_k. \quad (2.1.5)$$

El número de elementos distintos en una muestra aleatoria S con reemplazo es llamado **tamaño de muestra efectivo** y con probabilidad uno es menor o igual a N.

2.1.3 Soportes de muestreo

En los próximos capítulos empezará el tratamiento particular para estrategias de muestreo específicas; es decir, diseños de muestreo que se ajustan a ciertas situaciones y estimadores que mejoran la eficiencia de la estrategia. Sin embargo, antes de proseguir, es necesario que el lector entienda que las estrategias de muestreo se definen en términos del tipo de muestreo que se utiliza para la selección de muestras. En general, existen dos distinciones básicas.

1. **Tipo de muestreo:** selección de unidades con reemplazo o sin reemplazo.

2. **Tamaño de muestra:** tamaño de muestra fijo o aleatorio.

Como se verá en los capítulos posteriores, dependiendo de las anteriores condiciones, se define la estrategia de muestreo, el tratamiento teórico para la estimación de parámetros y el tipo de soporte. Esta sección trata específicamente sobre las diferentes formas que puede tomar el soporte de un diseño de muestreo dependiendo de las dos distinciones básicas. Para entrar en materia, es necesario enunciar las siguientes definiciones.

Definición 2.1.4. *Un* ***soporte*** *Q es un conjunto de muestras.*

Definición 2.1.5. *Un* ***soporte se llama simétrico*** *si para cualquier $s \in Q$, todas las permutaciones de s están también en Q.*

En los siguientes capítulos, a menos que se mencione lo contrario, el término **soporte** hará referencia a un **soporte simétrico**. Algunos soportes simétricos particulares son:

- El *soporte simétrico sin reemplazo* definido como

$$\mathcal{S} = \{0,1\}^N$$

  Nótese que

$$\#(\mathcal{S}) = 2^N$$

  Por ejemplo, si $N = 3$, entonces $\mathcal{S}$ queda definido por las siguientes muestras:

$$\mathcal{S} = \{(0,0,0)', (1,0,0)', (0,0,1)', (1,0,1)', (0,1,0)', (1,1,0)', (0,1,1)', (1,1,1)'\}$$

- El *soporte simétrico sin reemplazo de tamaño fijo* definido como

$$\mathcal{S}_n = \left\{\mathbf{s} \in \mathcal{S} | \sum_{k \in U} s_k = n\right\}$$

  Nótese que

$$\#(\mathcal{S}_n) = \binom{N}{n}$$

  Por ejemplo, si $N = 3$ y $n = 2$, entonces $\mathcal{S}_n$ queda definido por las siguientes muestras:

$$\mathcal{S}_n = \{(1,0,1)', (1,1,0)', (0,1,1)'\}$$

- El *soporte simétrico con reemplazo* definido como

$$\mathcal{R} = \mathbb{N}^N$$

  donde $\mathbb{N}$ es el conjunto de los números naturales. Nótese que este soporte es un conjunto contable pero infinito, por tanto

$$\#(\mathcal{R}) = \infty$$

- El *soporte simétrico con reemplazo de tamaño fijo* definido como

$$\mathcal{R}_m = \left\{\mathbf{s} \in \mathcal{R} | \sum_{k \in U} n_k = m\right\}$$

Nótese que

$$\#(\mathcal{R}_m) = \binom{N+m-1}{m}$$

Por ejemplo, si $N = 3$ y $m = 2$, entonces $\mathcal{R}_m$ queda definido por las siguientes muestras:

$$\mathcal{R}_m = \{(2,0,0)', (0,0,2)', (0,2,0)', (1,1,0)', (1,0,1)', (0,1,1)'\}$$

Tillé (2006) afirma que geométricamente cada vector $\mathbf{s}$ representa el vértice de un N-cubo. Además, se tiene el siguiente resultado:

Resultado 2.1.1. *Para los soportes definidos anteriormente, se tienen las siguientes propiedades:*

1. $\mathcal{S}, \mathcal{S}_n, \mathcal{R}, \mathcal{R}_m$ *son soportes simétricos.*
2. $\mathcal{S} \subset \mathcal{R}$.
3. *El conjunto* $\{\mathcal{S}_0, \mathcal{S}_1, \ldots, \mathcal{S}_N\}$ *es una partición de* $\mathcal{S}$.
4. *El conjunto* $\{\mathcal{R}_0, \mathcal{R}_1, \ldots, \mathcal{S}_N, \ldots\}$ *es una partición infinita de* $\mathcal{R}$.
5. $\mathcal{S} \subset \mathcal{R}$ *para todo* $n = 0, 1, \ldots, N$.

Muestras probabilísticas

No todas las muestras aleatorias son de tipo probabilístico. Una muestra (con o sin reemplazo) es de tipo probabilístico sí:

- Es posible construir (o al menos definir teóricamente) un soporte Q, tal que $Q = \{s_1, \ldots, s_q, \ldots, s_Q\}$, de todas las muestras posibles obtenidas por un método de selección. En donde s_q, $q = 1, \ldots, Q$, es una muestra perteneciente al soporte Q.
- Las probabilidades de selección que el proceso aleatorio le otorga a cada posible muestra perteneciente al soporte son conocidas de antemano a la selección de la muestra final.

Nótese que una muestra al azar no necesariamente es una muestra probabilística. En la mala práctica, algunos investigadores utilizan métodos aleatorios de inclusión de elementos sin disponer de un marco de muestreo y sin cumplir las dos condiciones anteriores; de esta manera, aunque los elementos sean escogidos de manera aleatoria o al azar, la muestra resultante no se puede catalogar como una muestra probabilística. Desde aquí en adelante, a menos que se diga lo contrario, el término muestra se refiere a una muestra probabilística. Algunos comentarios de interés son:

1. El universo U es finito.
2. La muestra probabilística s puede contener objetos repetidos. Esto sucede cuando el procedimiento de muestreo es con reemplazo.

3. La muestra s con repeticiones, puede tener un tamaño mayor al de la población.
4. La muestra s sin repeticiones, puede tener un tamaño máximo igual a N.
5. Si se presenta la ausencia del marco de muestreo es imposible realizar un procedimiento de muestreo probabilístico. Excepto cuando se realiza un censo.
6. Si la muestra seleccionada no es de tipo probabilístico, entonces no se puede construir ninguna estimación de tipo estadístico.
7. El estadístico deberá responder por los engaños o fraudes, que por ignorancia, mala fe o por la comodidad de mantener un empleo o negocio, para el cual no está capacitado, cometa contra clientes, ciudades y países que confían en la cifras resultantes de sus análisis.

Ejemplo 2.1.1. Suponga una población finita de tamaño $N = 5$, en donde los integrantes de la población están identificados cada uno con su nombre. La población la conforman los siguientes elementos:

Yves, Ken, Erik, Sharon, y Leslie,

En R se utiliza un vector de cadena de texto para indexar la población. Nótese que los elementos pertenecientes al vector son especificados mediante el uso de las comillas. En este caso los identificadores de cada elemento de la población, son asignados al objeto `U`.

```
> U  <-  c("Yves", "Ken", "Erik", "Sharon", "Leslie")
> U[1]
[1] "Yves"
> U[2]
[1] "Ken"
```

Para obtener el soporte Q, de todas las posibles muestras de tamaño $n = 2$ de esta población de tamaño $N = 5$, se utiliza la función `Support` del paquete `TeachingSampling`. Esta función contiene tres argumentos: el tamaño de la población `N`, el tamaño fijo de cada una de las posibles muestras `n` y, por último, una característica `y` que puede ser de tipo numérico o puede ser un conjunto de rótulos, la salida de la función será un conjunto de datos conteniendo todas las posibles muestras de tamaño fijo. Cuando el argumento `y` es distinto de `FALSE`, el resultado de la función será la característica poblacional para cada individuo. En el siguiente ejemplo se utiliza la función `Support(N,n,y=FALSE)` para obtener el conjunto de posibles muestras de tamaño dos de la población U, mientras que la función `Support(N,n,U)` arroja el conjunto de los rótulos en cada una de las 10 posibles muestras.

```
> N  <-  length(U)
> N
[1] 5
> n <- 2
```

```
> Support(N,n)
      [,1] [,2]
 [1,]    1    2
 [2,]    1    3
 [3,]    1    4
 [4,]    1    5
 [5,]    2    3
 [6,]    2    4
 [7,]    2    5
 [8,]    3    4
 [9,]    3    5
[10,]    4    5
```

```
> Support(N,n,U)
      [,1]     [,2]
 [1,] "Yves"   "Ken"
 [2,] "Yves"   "Erik"
 [3,] "Yves"   "Sharon"
 [4,] "Yves"   "Leslie"
 [5,] "Ken"    "Erik"
 [6,] "Ken"    "Sharon"
 [7,] "Ken"    "Leslie"
 [8,] "Erik"   "Sharon"
 [9,] "Erik"   "Leslie"
[10,] "Sharon" "Leslie"
```

2.1.4 Diseño de muestreo

Definición 2.1.6. *Un **diseño de muestreo** $p(\cdot)$ es una distribución de probabilidad multivariante definida sobre un soporte Q; es decir, $p(\cdot)$ es una función que va desde Q[2] hasta $(0,1]$ tal que $p(s)>0$ para todo $s \in Q$ y*

$$\sum_{s \in Q} p(s) = 1 \tag{2.1.6}$$

Dado el soporte Q, un **diseño de muestreo** es una función $p(\cdot)$, tal que $p(s)$ arroja la probabilidad de selección de la muestra realizada s bajo un esquema de selección particular. En otras palabras, si S es una muestra aleatoria que toma el valor s con probabilidad $p(s)$, tal que

$$Pr(S = s) = p(s) \qquad \text{para todo } s \in Q. \tag{2.1.7}$$

Entonces $p(\cdot)$ es llamada diseño de muestreo.

El diseño muestreo, es una función que va desde el soporte Q hasta el intervalo $]0,1]$. Por ser una distribución de probabilidad se tiene que $p(\cdot)$ cumple que

1. $p(s) \geq 0$ para todo $s \in Q$
2. $\sum_{s \in Q} p(s) = 1$

Nótese que el diseño de muestreo no se refiere a un algoritmo o procedimiento que permite la selección de muestras. Dado un diseño de muestreo, el trabajo del estadístico consiste en encontrar un algoritmo que permita la selección de muestras cuya probabilidad de selección corresponda a la probabilidad inducida por el diseño de muestreo. Para la realización de inferencias acerca de los parámetros de interés, el diseño de muestreo juega un papel muy importante porque las propiedades estadísticas (esperanza, varianza y otros) de las cantidades aleatorias que se calculan basadas en una muestra están determinadas por éste.

Dado un soporte Q, un diseño de muestreo puede ser:

[2]Nótese que Q es el espacio muestral cuyos elementos son vectores.

- **Sin reemplazo** si todas las posibles muestras en Q son sin reemplazo.
- **Con reemplazo** si todas las posibles muestras en Q son con reemplazo.
- **De tamaño fijo** si todas las posibles muestras en Q tienen el mismo tamaño de muestra $n(S) = n$.

Cassel, Särndal & Wretman (1976*a*) explican que la posibilidad de identificar cada una de todas las posibles muestras que pertenecen al soporte Q es un factor crucial que permite:

- designar un conjunto de muestras a las cuales se les asigna una probabilidad positiva de selección y
- distribuir la totalidad de la masa de probabilidad entre los miembros de Q.

El rasgo más importante del muestreo probabilístico es que permite conocer, por lo menos teóricamente, la probabilidad de selección de todas las posibles muestras en el soporte Q. Sin embargo, un diseño de muestreo también deja conocer la probabilidad de inclusión del elemento k en la muestra S.

Algoritmo de selección

Un diseño de muestreo es una distribución de probabilidad sobre un soporte Q; pero, de ninguna manera, es un procedimiento que selecciona la muestra per se.

Definición 2.1.7. *Un* ***algoritmo de selección*** *es un procedimiento usado para seleccionar una muestra probabilística.*

Tillé (2006) afirma que una forma de seleccionar una muestra es listar todas las posibles muestras, generar una variable aleatoria con distribución uniforme en el intervalo $[0, 1]$ para luego hacer la correspondiente selección. A este tipo de algoritmos que listan todas las posibles muestras se les conoce con el nombre de **algoritmos de selección enumerativos**; sin embargo, este tipo de algoritmos son ineficientes computacionalmente y sólo son posibles de implementar cuando el diseño de muestreo es conocido y el tamaño poblacional N es pequeño. A lo largo del libro se incluirán diversos algoritmos de selección específicos para cada diseño de muestreo que permitan la selección de una muestra probabilística.

2.1.5 Probabilidad de inclusión

La inclusión del elemento k-ésimo en una muestra s particular es un evento aleatorio definido por la función indicadora $I_k(s)$, que está dada por

$$I_k(s) = \begin{cases} 1 & \text{si } k \in s \\ 0 & \text{si } k \notin s. \end{cases} \tag{2.1.8}$$

Nótese que la función $I_k(s)$ es una función de la variable aleatoria S. Para acortar la notación escribiremos $I_k = I_k(s)$, entendiéndose que I_k es la función indicadora

para el elemento k-ésimo. Bajo un diseño de muestreo $p(\cdot)$, una **probabilidad de inclusión** es asignada a cada elemento de la población para indicar la probabilidad de que el elemento pertenezca a la muestra. Para el elemento k-ésimo de la población, la probabilidad de inclusión se denota como π_k y se conoce como la probabilidad de inclusión de **primer orden** y está dada por

$$\pi_k = Pr(k \in S) = Pr(I_k = 1) = \sum_{s \ni k} p(s). \tag{2.1.9}$$

En donde el subíndice $s \ni k$ se refiere a la suma sobre todas las muestras que contienen al elemento k-ésimo. Nótese que de la anterior definición para que una muestra sea considerada probabilística, entonces todos los elementos en la población deben tener probabilidad de inclusión estrictamente mayor a cero.

Definición 2.1.8. *La **esperanza de una muestra** aleatoria, en el sentido de las definiciones 2.1.2. y 2.1.3., está dada por*

$$\boldsymbol{\mu} = E(\mathbf{s}) = \sum_{\mathbf{s} \in Q} p(\mathbf{s})\mathbf{s} \tag{2.1.10}$$

Si el diseño muestral es sin reemplazo, entonces $\boldsymbol{\mu} = \boldsymbol{\pi}$, donde $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)'$ es el vector de probabilidades de inclusión inducido por el diseño de muestreo. El siguiente resultado provee una manera sencilla para computar y realizar el cálculo de las N probabilidades de inclusión.

Resultado 2.1.2. *Dado un soporte Q, la probabilidad de inclusión π_k es la probabilidad de que el elemento k-ésimo pertenezca a la muestra aleatoria S y se puede escribir de la siguiente manera:*

$$\pi_k = E(I_k(S)) = \sum_{s \in Q} I_k(s) p(s) \tag{2.1.11}$$

Prueba. $I_k(S)$ es una función de la muestra aleatoria S, la demostración se sigue de la definición de la esperanza de una función de una variable aleatoria. Por otro lado, $I_k(S)$ sólo puede tomar dos valores 1 y 0, luego

$$\begin{aligned} E(I_k(S)) &= (1)Pr(I_k(S) = 1) + (0)Pr(I_k(S) = 0) \\ &= Pr(I_k(S) = 1) = Pr(k \in S) = \pi_k \end{aligned}$$

■

Análogamente, π_{kl} se conoce como la probabilidad de inclusión de **segundo orden** y denota la probabilidad de que los elementos k y l pertenezcan a la muestra, ésta se denota como π_{kl} y está dada por

$$\pi_{kl} = Pr(k \in S \text{ y } l \in S) = Pr(I_k I_l = 1) = \sum_{s \ni \ k \text{ y } l} p(s). \tag{2.1.12}$$

En donde el subíndice $s \ni \ k$ y l se refiere a la suma sobre todas las muestras que contienen a los elementos k-ésimo y l-ésimo.

Ejemplo 2.1.2. Considere el siguiente diseño de muestreo $p(\cdot)$ tal que asigna las siguientes probabilidades de selección a cada una de las 10 posibles muestras de tamaño 2 del soporte Q de la población U.

```
> p <- c(0.13,0.2,0.15,0.1,0.15,0.04,0.02,0.06,0.07,0.08)
> p
 [1] 0.13 0.15 0.15 0.10 0.15 0.04 0.02 0.06 0.07 0.08
```

Es decir, la primera muestra tiene una probabilidad de selección de 0.13, la segunda muestra tiene una probabilidad de selección de 0.15, y así sucesivamente hasta la décima cuya probabilidad de selección es de 0.08. Con las siguientes instrucciones verificamos que las propiedades de diseño muestral sean satisfechas.

```
> sum(p)
[1] 1

> p<0
FALSE FALSE FALSE FALSE FALSE
FALSE FALSE FALSE FALSE FALSE
```

Mediante el uso de la función `Ik` del paquete `TeachingSampling`, es posible crear las $N = 5$ funciones indicadoras de los elementos pertenecientes a la población para cada una de las 10 posibles muestras de tamaño fijo y sin reemplazo. Esta función contiene dos argumentos: el tamaño de la población `N`, el tamaño fijo de cada una de las posibles muestras `n`. Una tabla de datos es creada a partir de los rótulos, la probabilidad de selección y las 5 funciones indicadoras de las posibles muestras contenidas en el soporte Q.

```
> Ind <- Ik(N,n)
> Q <- Support(N,n,U)

> data.frame(Q,p,Ind)
       V1     V2       p    I1 I2 I3 I4 I5
1    Yves    Ken     0.13    1  1  0  0  0
2    Yves   Erik     0.20    1  0  1  0  0
3    Yves Sharon     0.15    1  0  0  1  0
4    Yves Leslie     0.10    1  0  0  0  1
5     Ken   Erik     0.15    0  1  1  0  0
6     Ken Sharon     0.04    0  1  0  1  0
7     Ken Leslie     0.02    0  1  0  0  1
8    Erik Sharon     0.06    0  0  1  1  0
9    Erik Leslie     0.07    0  0  1  0  1
10 Sharon Leslie     0.08    0  0  0  1  1
```

Una vez son calculadas las variables indicadoras para cada elemento y en cada posible muestra, el cálculo de las probabilidades de inclusión se hace muy sencillo al multiplicar las probabilidades de selección con cada una de las variables indicadoras. El resultado se suma por columnas y la salida es un vector de tamaño $N = 5$ de probabilidades de inclusión.

```
> multip <- p*Ind
> colSums(multip)
[1] 0.58 0.34 0.48 0.33 0.27
```

La función `Pik` del paquete `TeachingSampling` arroja el vector de probabilidades de inclusión para todos los elementos de la población. Ésta tiene dos argumentos: un vector `p` de probabilidades de selección de todas las posibles muestras y una matriz `Ind` de N variables indicadoras. Nótese que la suma de probabilidades de inclusión es el tamaño de muestra esperado, en este caso igual a 2.

```
> pik <- Pik(p, Ind)
> pik
[1] 0.58 0.34 0.48 0.33 0.27
```

Luego, el elemento de la población que tiene una mayor probabilidad de ser incluido es **Yves**, mientras que el elemento con una menor probabilidad de inclusión es **Sharon**. Por otra parte, haciendo uso de la función `Pikl` del paquete `TeachingSampling` es posible calcular la matriz de probabilidades de inclusión de segundo orden para el diseño `p` en cuestión. Esta función sólo tiene tres argumentos: `N`, el tamaño de la población, `n`, el tamaño de muestra fijo y `p`, el diseño de muestreo utilizado. La salida de esta función es una matriz cuadrada y simétrica de tamaño $N \times N$ cuyas entradas corresponden a las probabilidades de inclusión de segundo orden. Para este caso particular tenemos que la función se ejecuta de la siguiente manera.

```
> pikl <- Pikl(N, n, p)
> pikl
     [,1] [,2] [,3] [,4] [,5]
[1,] 0.58 0.13 0.20 0.15 0.10
[2,] 0.13 0.34 0.15 0.04 0.02
[3,] 0.20 0.15 0.48 0.06 0.07
[4,] 0.15 0.04 0.06 0.33 0.08
[5,] 0.10 0.02 0.07 0.08 0.27
```

Nótese que, bajo este diseño de muestreo, **Yves** y **Erik** corresponden al par de elementos que tienen la más alta probabilidad de inclusión.

2.1.6 Característica de interés y parámetros de interés

El propósito de cualquier estudio por muestreo es estudiar una **característica** de interés y que se encuentra asociada a cada unidad de la población. Es decir, la característica de interés toma el valor y_k para la unidad k. Es importante notar que los y_ks no se consideran variables aleatorias sino cantidades fijas, por tanto la notación de éstas se hace con un letra minúscula y. El objetivo de la investigación por muestreo es estimar una función de interés T, llamada **parámetro**, de la característica de interés en la población.

$$T = f\{y_1, \ldots, y_k, \ldots, y_N\}.$$

Algunos de los parámetros de interés más comunes son:

1. El total poblacional,
$$t_y = \sum_{k \in U} y_k \tag{2.1.13}$$
2. La media poblacional,
$$\bar{y}_U = \frac{\sum_{k \in U} y_k}{N} = \frac{t_y}{N} \tag{2.1.14}$$
3. La varianza poblacional,
$$S^2_{yU} = \frac{\sum_{k \in U}(y_k - \bar{y}_U)^2}{N-1} \tag{2.1.15}$$

Existen otros parámetros de interés como la mediana poblacional, los percentiles poblaciones, la razón entre dos totales poblacionales o, como se mencionó anteriormente, el tamaño de una población, en cuyo caso estaríamos interesados en N. Entre otros, algunos ejemplos de investigaciones por muestreo interesadas en los anteriores parámetros son:

- Total de personas que pertenecen a la fuerza laboral.
- Porcentaje de personas que usarían un producto.

Obviamente, estas cantidades poblacionales son desconocidas y ésta es la razón por la que se requiere realizar una investigación por muestreo, porque mediante ésta se pueden estimar estos parámetros poblacionales a partir de una muestra seleccionada.

Ejemplo 2.1.3. Suponga que en nuestra población de ejemplo se quiere estimar el total de la variable y. El valor para cada uno de los elementos de la población es el siguiente:

```
> y <- c(32,34,46,89,35)
> y
[1] 32 34 46 89 35
```

La función `data.frame` crea el conjunto de datos conteniendo los nombres (rótulos) y el valor de la característica de interés para cada elemento de la población

```
> data.frame(U,y)
        U       y
1       Yves    32
2       Ken     34
3       Erik    46
4       Sharon  89
5       Leslie  35
```

Algunos parámetros poblacionales de interés de la característica y son, el total poblacional y la media dados por t_y y $\bar{y}_U$, respectivamente.

```
> ty  <-  sum(y)
> ty
[1] 236
> ybar  <-  ty/N
> ybar
[1] 47.2
```

2.1.7 Estadística y estimador

Una **estadística** es una función G (que toma valores reales) de la muestra aleatoria S y sólo depende de los elementos pertenecientes a S. Cuando una estadística se usa para estimar un parámetro se dice **estimador** y las realizaciones del estimador en una muestra seleccionada s se dicen **estimaciones**.

Siendo G una estadística, sus propiedades estadísticas están determinadas por el diseño de muestreo. Es decir, dada la probabilidad de selección de cada muestra $s \in Q$, la esperanza, la varianza y otras propiedades de interés están definidas a partir de $p(s)$.

La **esperanza** de una estadística G es

$$E(G) = \sum_{s \in Q} p(s)G(s). \tag{2.1.16}$$

La **varianza** de la estadística G está definida como

$$Var(G) = E[G - E(G)]^2 \tag{2.1.17}$$
$$= \sum_{s \in Q} p(s)[G(s) - E(G)]^2. \tag{2.1.18}$$

Donde $G(s)$ es el valor real que toma la estadística G en la muestra seleccionada (realizada) s y Q es el soporte inducido por el diseño muestral. Nótese que las propiedades de las estadísticas y, por consiguiente, de los estimadores, están definidas con sumas porque el diseño de muestreo induce una distribución de probabilidad discreta sobre todas las posibles muestras s pertenecientes al soporte Q.

La estadística I_k

La cantidad I_k dada por (2.1.8) es una estadística que toma valores aleatoriamente dependiendo del diseño de muestreo utilizado.

Resultado 2.1.3. *Las propiedades más importantes de esta estadística son:*

- $E(I_k) = \pi_k$
- $Var(I_k) = \pi_k(1 - \pi_k)$
- $Cov(I_k, I_l) = \pi_{kl} - \pi_k\pi_l$ *para todo* $k \neq l$

Prueba. Por el resultado 2.1.2., la primera propiedad se tiene de inmediato, ahora de la definición de varianza se tiene

$$\begin{aligned}Var(I_k(S)) &= E[I_k(S) - E(I_k(S))]^2\\ &= Pr(I_k(S)=1)[1-\pi_k]^2 + Pr(I_k(S)=0)[0-\pi_k]^2\\ &= \pi_k(1-\pi_k)\end{aligned}$$

y finalmente, de la definición de covarianza se tiene

$$\begin{aligned}Cov(I_k(S), I_l(S)) &= E[I_k(S)I_l(S)] - E[I_k(S)]E[I_k(S)]\\ &= (1)Pr(I_k(S)I_l(S)=1) + (0)Pr(I_k(S)I_l(S)=0) - \pi_k\pi_l\\ &= \pi_{kl} - \pi_k\pi_l\end{aligned}$$

■

A la covarianza de las estadísticas indicadoras para los elementos k y l, $Cov(I_k, I_l)$, se le conoce como Δ_{kl}. Esta cantidad, dependiendo del diseño, puede tomar valores positivos, negativos o incluso nulos.

La estadística $n(S)$ o tamaño de muestra

Como ya se vio, el tamaño de muestra es una cantidad aleatoria, dependiendo del diseño. Nótese que este valor puede ser expresado como función de las estadísticas de inclusión.

$$n(S) = \sum_U I_k. \tag{2.1.19}$$

Resultado 2.1.4. *Algunas propiedades de interés son:*

- $E(n(S)) = \sum_U \pi_k$
- $Var(n(S)) = \sum_U \pi_k - (\sum_U \pi_k)^2 + \sum\sum_{k\neq l} \pi_{kl}$.

Prueba. Para la primera propiedad, se tiene que

$$E[n(S)] = E\left[\sum_U I_k\right] = \sum_U E[I_k] = \sum_U \pi_k$$

Recordando que las propiedades de la varianza de una suma se tiene

$$\begin{aligned}Var[n(S)] &= Var\left[\sum_U I_k\right]\\ &= \sum_U Var[I_k] + \sum\sum_{k\neq l} Cov[I_k, I_l]\\ &= \sum_U \pi_k - \sum_U \pi_k^2 - \sum\sum_{k\neq l}\pi_k\pi_l + \sum\sum_{k\neq l}\pi_{kl}\\ &= \sum_U \pi_k - \left(\sum_U \pi_k\right)^2 + \sum\sum_{k\neq l}\pi_{kl}\end{aligned}$$

■

Además, cuando la variación del tamaño de muestra es nula porque se ha decidido utilizar un diseño de tamaño muestral fijo, se tienen las siguientes propiedades.

Resultado 2.1.5. *Si el diseño de muestreo es de tamaño fijo e igual a n,*

- $E(n(S)) = \sum_U \pi_k = n$
- $\sum_U \pi_{kl} = n\pi_l$
- $\sum_U \Delta_{kl} = 0$
- $\pi_k(1-\pi_k) = \sum_{l\neq k}(\pi_k\pi_l - \pi_{kl})$

Prueba. La primera propiedad se tiene recordando que la esperanza de una constante es ella misma. Nótese que $\pi_{kl} = E[I_k(S)I_l(S)]$, así

$$\begin{aligned}\sum_{l\in U}\pi_{kl} = \sum_{l\in U} E[I_k(S)I_l(S)] &= \sum_{l\in U}\sum_{s\in Q} p(s)I_k(s)I_l(s)\\ &= \sum_{s\in Q} p(s)I_k(s)\sum_{l\in U} I_l(s)\\ &= n(S)\sum_{s\in Q} p(s)I_k(s) = n\pi_k\end{aligned}$$

La tercera propiedad se tiene pues

$$\begin{aligned}\sum_U \Delta_{kl} &= \sum_U(\pi_{kl} - \pi_k\pi_l)\\ &= \sum_U \pi_{kl} - \pi_k\sum_U \pi_l\\ &= n\pi_k - n\pi_k = 0\end{aligned}$$

Para demostrar la última propiedad es necesario redefinir el tamaño de muestra, de tal manera que $n = \sum_{l\neq k} I_l(S) + I_k(S)$. Luego,

$$\begin{aligned}\pi_k(1-\pi_k) &= Var(I_k(S))\\ &= Cov(I_k(S), I_k(S))\\ &= Cov\left(I_k(S), n - \sum_{l\neq k} I_l(S)\right)\\ &= -\sum_{l\neq k} Cov(I_k(S), I_l(S))\\ &= \sum_{l\neq k}(\pi_k\pi_l - \pi_{kl})\end{aligned}$$

■

Ejemplo 2.1.4. Continuando con el desarrollo del ejemplo 2.1.3, ahora utilizaremos el vector de probabilidades de inclusión y la matriz de probabilidades de segundo orden para verificar los resultados 2.1.4 y 2.1.5. En primer lugar, nótese que la esperanza del tamaño de muestra, que corresponde a 2 pues el diseño es de tamaño fijo, se obtiene de la siguiente manera.

```
> A <- sum(pik)
> A
[1] 2
```

Ahora, el cuadrado de la suma de las probabilidades de inclusión se obtiene así

```
> B <- (sum(pik))^2
> B
[1] 4
```

Y la suma de los elementos distintos de la matriz de probabilidades de inclusión de segundo orden es

```
> C <- sum(pikl)-sum(diag(pikl))
> C
[1] 2
```

Para comprobar la segunda parte del resultado 2.1.4. basta realizar la siguiente operación `A-B+C`. Esta suma es nula y efectivamente corresponde a la varianza del tamaño de muestra en este diseño de muestreo; como, en este caso particular, el tamaño de muestra siempre fue fijo e igual a 2, la varianza debe ser cero.

El siguiente paso de este ejemplo consiste en la verificación de la segunda parte del resultado 2.1.5. En resumidas cuentas, este apartado dice que la suma por filas (o columnas) de la matriz de probabilidades de inclusión de segundo orden debe corresponder exactamente a la multiplicación del tamaño de muestra y el vector de probabilidades de inclusión de primer orden. Lo anterior se corrobora fácilmente por medio del siguiente código.

```
> n*pik
     [,1] [,2] [,3] [,4] [,5]
[1,] 1.16 0.68 0.96 0.66 0.54
> colSums(pikl)
[1] 1.16 0.68 0.96 0.66 0.54
> rowSums(pikl)
[1] 1.16 0.68 0.96 0.66 0.54
```

Nótese que la suma por filas y por columnas coincide perfectamente con $n \times \pi_k$ para todo $k \in U$. Por otro lado, verificaremos la tercera propiedad que afirma que la suma por filas (o columnnas) de la matriz de varianzas-covarianzas de las variables indicadoras de membresía muestral debe dar como resultado un vector de ceros de tamaño cinco. Para esto, se utiliza la función `Deltakl` del paquete `TeachingSampling`. Esta función tiene tres argumentos: `N`, el tamaño de la población, `n`, el tamaño de muestra fijo y `p`, el diseño de muestreo utilizado. La salida de esta función corresponde a una matriz cuadrada y simétrica de tamaño $N \times N$ cuyas entradas corresponden a las varianzas-covarianzas de las variables indicadoras de membresía muestral. Para este ejemplo, la implementación del siguiente código permite obtener la matriz buscada y la verificación del resultado.

```
> Delta <- Deltakl(N, n, p)
```

```
> Delta

         [,1]    [,2]    [,3]    [,4]    [,5]
[1,]  0.2436 -0.0672 -0.0784 -0.0414 -0.0566
[2,] -0.0672  0.2244 -0.0132 -0.0722 -0.0718
[3,] -0.0784 -0.0132  0.2496 -0.0984 -0.0596
[4,] -0.0414 -0.0722 -0.0984  0.2211 -0.0091
[5,] -0.0566 -0.0718 -0.0596 -0.0091  0.1971

> rowSums(Delta)
[1] 0   0   0   0   0
> colSums(Delta)
[1] 0   0   0   0   0
```

De esta manera la suma por filas (o columnas) de la matriz de varianzas-covarianzas de las variables indicadoras de membresía muestral es cero en cada columna (o fila).

Cuando una estadística se construye con la intención de estimar un parámetro, recibe el nombre de **estimador**. Así, las propiedades más comúnmente utilizadas de un estimador $\hat{T}$ de un parámetro de interés T son el sesgo, definido por

$$B(\hat{T}) = E(\hat{T}) - T \tag{2.1.20}$$

y el error cuadrático medio, dado por

$$ECM(\hat{T}) = E[\hat{T} - T]^2 \tag{2.1.21}$$
$$= Var(\hat{T}) + B^2(\hat{T}). \tag{2.1.22}$$

Si el sesgo de un estimador es nulo se dice que el estimador es **insesgado** y cuando esto ocurre el error cuadrático medio se convierte en la varianza del estimador.

Särndal, Swensson & Wretman (1992) afirman que el objetivo en un estudio por muestreo es estimar uno a más parámetros poblacionales. Las decisiones más importantes a la hora de abordar un problema de estimación por muestreo son

- La escogencia de un diseño de muestreo y un algoritmo de selección que permita implementar el diseño.
- La elección de una fórmula matemática o estimador que calcule una estimación del parámetro de interés en la muestra seleccionada.

Las anteriores no son decisiones independientes. Es decir, la escogencia de un estimador dependerá, usualmente, del diseño de muestreo utilizado.

Definición 2.1.9. *Siendo $\hat{T}$ un estimador de un parámetro T y $p(\cdot)$ un diseño de muestreo definido sobre un soporte Q, se define una* ***estrategia de muestreo*** *como la dupla $(p(\cdot), \hat{T})$.*

Este libro, como su nombre lo indica, está enfocado en la búsqueda de la mejor combinación de diseño de muestreo y estimador; este problema ha sido considerado a través del desarrollo de la teoría de muestreo. La escogencia de la estrategia de muestreo se lleva a cabo en dos etapas, a saber: **Etapa de diseño**, refiriéndose al periodo durante el cual se decide el diseño de muestreo a utilizar junto con el algoritmo de muestreo que permita la selección de la muestra y finalmente se selecciona la muestra probabilística. Una vez que la información es recogida y grabada entra la **Etapa de estimación** en donde se calculan las estimaciones para la característica de interés utilizando el estimador propio de la estrategia de muestreo escogida.

2.2 Estimadores de muestreo

Cada elemento perteneciente a la población tiene una característica de interés asociada y. Para el elemento k-ésimo el valor que toma esta característica de interés es y_k. El objetivo de la investigación por muestreo es estimar un parámetro T que resulta de interés. El objetivo del estadístico es poder inferir acerca de T con base en una muestra s. Un indicador de la precisión de un estimador está dado por el **coeficiente de variación estimado** dado por

$$cve(\hat{T}) = \frac{\sqrt{\widehat{Var}(\hat{T})}}{\hat{T}} \tag{2.2.1}$$

donde $\widehat{Var}(\hat{T})$ es el estimador de la varianza basado en la muestra seleccionada s. El coeficiente de variación estimado es una medida comúnmente usada para expresar el error cometido al seleccionar una muestra y ni utilizar a toda la población en la medición de la variable de interés. Si se realizara un censo y el estimador reprodujera el parámetro poblacional, entonces $\widehat{Var}(\hat{T})$ sería nula y, por lo tanto, el *cve* también sería nulo.

A continuación, se revisan algunos de los estimados más utilizados en la historia del muestreo. A medida que se avance en la lectura del libro, nuevos estimadores surgirán y, por consiguiente, nuevas estrategias de muestreo que permiten llegar a resultados con una precisión casi clínica. La mayoría de los estimadores presentados en este libro son estimadores de totales o de funciones de totales.

2.2.1 El estimador de Horvitz-Thompson

Estimador del total poblacional

Narain (1951) descubrió este estimador, aunque su artículo fue editado y publicado por una revista india de poca rotación. Más adelante Horvitz & Thompson (1952) publicaron similares resultados en la revista más importante de estadística en ese tiempo, JASA (Journal of the American Statistical Society). Desde entonces, este estimador se conoce como el estimador de Horvitz-Thompson o estimador π, aunque rigurosamente debería ser llamado estimador de Narain-Horvitz-Thompson . En este libro seguiremos la notación internacional y clásica.

Para un universo U, se quiere estimar el total poblacional t_y de la característica de interés y dado por (2.1.13). Se define el estimador de Horvitz-Thompson(HT) para t_y como:

$$\hat{t}_{y,\pi} = \sum_S \frac{y_k}{\pi_k} = \sum_S d_k y_k \tag{2.2.2}$$

Donde π_k es la probabilidad de inclusión para el k-ésimo elemento, y d_k es conocido como **factor de expansión** y corresponde al inverso de la probabilidad de inclusión. Nótese que el estimador de Horvitz-Thompson es aleatorio porque está construido con base en una suma sobre la muestra aleatoria S. La motivación detrás de este estimador, como Brewer (2002) lo indica, descansa en el **principio de representatividad** que afirma que cada elemento incluido en una muestra se representa a sí mismo y a un grupo de unidades que no pertenecen a la muestra seleccionada, cuyas características son cercanas a las del elemento incluido en la muestra. El factor de expansión no es otra cosa que el número de elementos menos uno de la población (no incluidos en la muestra) representados por el elemento incluido.

Resultado 2.2.1. *Si todas las probabilidades de inclusión de primer orden son mayores a cero ($\pi_k > 0$ para todo k), el estimador de Horvitz-Thompson es insesgado para el total poblacional. Por tanto, se tiene que*

$$E(\hat{t}_{y,\pi}) = t_y \tag{2.2.3}$$

Prueba. Reescribiendo el estimador de Horvitz-Thompson como $\hat{t}_{y,\pi} = \sum_S I_k(S)\frac{y_k}{\pi_k}$, se tiene

$$E(\hat{t}_{y,\pi}) = E\left(\sum_U I_k(S)\frac{y_k}{\pi_k}\right) = \sum_U \frac{y_k}{\pi_k} E\left(I_k(S)\right) = \sum_U \pi_k \frac{y_k}{\pi_k} = t_y$$

■

Si el diseño de muestreo es tal que las probabilidades de inclusión de primer orden conservan una buena correlación positiva con la medición de la característica de interés; en otras palabras, si $\pi_k \propto y_k$, el estimador de Horvitz-Thompson se reduce a una constante, por lo tanto tendrá varianza nula. En la práctica, una estrategia de muestreo óptima (Cassel, Särndal & Wretman 1976*a*) es aquella que utiliza el estimador de Horvitz-Thompson junto con un diseño de muestreo que induzca una buena correlación entre el vector de probabilidades de inclusión y el vector de valores de la característica de interés. Sin embargo, en encuestas multi-propósito, en donde se quiere estimar parámetros para varias características de interés entre las cuales no hay una buena correlación, al utilizar el estimador de Horvitz-Thompson es difícil evadir la débil, e incluso negativa, correlación que existe entre las características de interés y el vector de probabilidades de inclusión. Sin embargo, al incluir información auxiliar en la construcción del estimador se puede palear este hecho.

Varianza del estimador de Horvitz-Thompson

Resultado 2.2.2. *La varianza del estimador de Horvitz-Thompson está dada por la siguiente expresión*

$$Var_1(\hat{t}_{y,\pi}) = \sum\sum_U \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}. \tag{2.2.4}$$

Prueba. De la definición de varianza, se obtiene lo siguiente

$$\begin{aligned}
Var_1(\hat{t}_{y,\pi}) &= Var\left(\sum_U I_k(S)\frac{y_k}{\pi_k}\right)\\
&= \sum_U \frac{y_k^2}{\pi_k^2} Var(I_k(S)) + \sum\sum_{k\neq l} \frac{y_k}{\pi_k}\frac{y_l}{\pi_l} Cov(I_k(S), I_l(S))\\
&= \sum_U \frac{y_k^2}{\pi_k^2}(\pi_k - \pi_k^2) + \sum\sum_{k\neq l} \frac{y_k}{\pi_k}\frac{y_l}{\pi_l}(\pi_{kl} - \pi_k\pi_l)\\
&= \sum\sum_U \frac{y_k}{\pi_k}\frac{y_l}{\pi_l}(\pi_{kl} - \pi_k\pi_l)\\
&= \sum\sum_U \Delta_{kl}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}
\end{aligned}$$

■

Sen (1953) y Yates & Grundy (1953) dedujeron el siguiente resultado cuando el diseño de muestreo es de tamaño fijo.

Resultado 2.2.3. *Si el diseño $p(\cdot)$ es de tamaño de muestra fijo, entonces, la varianza del estimador de Horvitz-Thompson se escribe como*

$$Var_2(\hat{t}_{y,\pi}) = -\frac{1}{2}\sum\sum_U \Delta_{kl}\left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2 \tag{2.2.5}$$

Prueba. Utilizando las propiedades del resultado 2.1.5, se tiene que

$$\begin{aligned}
Var_2(\hat{t}_{y,\pi}) &= -\frac{1}{2}\sum\sum_U \Delta_{kl}\left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2\\
&= -\frac{1}{2}\sum\sum_U \Delta_{kl}\left(\frac{y_k^2}{\pi_k^2} + \frac{y_l^2}{\pi_l^2} - 2\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}\right)\\
&= -\frac{1}{2}\left[\sum\sum_U \Delta_{kl}\frac{y_k^2}{\pi_k^2} + \sum\sum_U \Delta_{kl}\frac{y_l^2}{\pi_l^2} - 2\sum\sum_U \Delta_{kl}\frac{y_l}{\pi_k}\frac{y_k}{\pi_l}\right]\\
&= -\frac{1}{2}\left[2\sum\sum_U \Delta_{kl}\frac{y_k^2}{\pi_k^2} - 2\sum\sum_U \Delta_{kl}\frac{y_l}{\pi_k}\frac{y_k}{\pi_l}\right]\\
&= -\sum_U \frac{y_k^2}{\pi_k^2}\sum_U \Delta_{kl} + \sum\sum_U \Delta_{kl}\frac{y_l}{\pi_k}\frac{y_k}{\pi_l}\\
&= \sum\sum_U \Delta_{kl}\frac{y_l}{\pi_k}\frac{y_k}{\pi_l} = Var_1(\hat{t}_{y,\pi})
\end{aligned}$$

puesto que $\sum_U \Delta_{kl} = 0$ para diseños de tamaño fijo. Por lo tanto, en los casos de diseños de muestreo con tamaño fijo, la varianza del estimador de Horvitz-Thompson puede calcularse por medio de $Var_2(\hat{t}_{y,\pi})$. ■

Estimación de la varianza

Es posible construir dos estimadores insesgados para las expresiones (2.2.4) y (2.2.5). Para esto, se requiere que todas las probabilidades de inclusión de segundo orden sean estrictamente positivas ($\pi_{kl} > 0$ para todo k). Con el anterior supuesto, se tienen los siguientes resultados.

Resultado 2.2.4. *Un estimador insesgado para la expresión (2.2.4) está dada por*

$$\widehat{Var}_1(\hat{t}_{y,\pi}) = \sum\sum_S \frac{\Delta_{kl}}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \tag{2.2.6}$$

Resultado 2.2.5. *Si el diseño es de tamaño de muestra fijo, un estimador insesgado para la expresión (2.2.5) está dado por*

$$\widehat{Var}_2(\hat{t}_{y,\pi}) = -\frac{1}{2}\sum\sum_S \frac{\Delta_{kl}}{\pi_{kl}} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2 \tag{2.2.7}$$

Prueba. Los anteriores resultados son inmediatos al reescribir los estimadores $\widehat{Var}_1(\hat{t}_{y,\pi})$ y $\widehat{Var}_2(\hat{t}_{y,\pi})$ en términos de U y multiplicar por el producto de las funciones indicadoras $I_k(S)I_l(S)$. Al aplicar la esperanza se tiene que $E[I_k(S)I_l(S)] = \pi_{kl}$ y con esto se tiene la demostración. ■

Bautista (1998) resalta los tres siguientes comentarios importantes acerca de las estimaciones arrojadas por anteriores expresiones.

1. Si las probabilidades de inclusión de segundo orden son mayores que cero para todos los elementos en la muestra, pero no para los restantes elementos que no fueron incluidos en la muestra, no se puede garantizar el insesgamiento de las anteriores expresiones.

2. Es posible que las estimaciones de la varianza arrojen resultados negativos, que no pueden ser utilizados ni interpretados. Para evitar esta situación, es necesario garantizar que la covarianza entre las estadísticas de inclusión para cada par de elementos en la población sea negativa ($\Delta_{kl} < 0 \ \forall \ k \neq l$).

3. No necesariamente las estimaciones arrojadas por las anteriores expresiones coinciden en todos los casos.

Por su parte, Tillé (2006) agrega que en la práctica, la utilización de las expresiones de los estimadores de la varianza es muy difícil de implementar pues la doble suma hace que el proceso de cálculo computacional sea muy largo e ineficiente. Por lo tanto, para cada diseño de muestreo que se utilice, se deben crear expresiones que pueden ser simplificadas o en algunos casos se deben utilizar aproximaciones.

Intervalo de confianza para el estimador de Horvitz-Thompson

Hájek (1960) demuestra la convergencia asintótica del estimador de Horvitz-Thompson a una distribución normal. Cuando el tamaño de muestra es suficientemente grande (que dependiendo del comportamiento de la población puede bastar con algunas docenas de individuos), se puede construir un intervalo de confianza de nivel $(1-\alpha)$ para el total poblacional t_y de acuerdo con:

$$IC(1-\alpha) = \left[\hat{t}_{y,\pi} - z_{1-\alpha/2}\sqrt{Var(\hat{t}_{y,\pi})}, \hat{t}_{y,\pi} + z_{1-\alpha/2}\sqrt{Var(\hat{t}_{y,\pi})}\right] \qquad (2.2.8)$$

donde $z_{1-\alpha/2}$ se refiere al cuantil $(1-\alpha/2)$ de una variable aleatoria con distribución normal estándar. Nótese que

$$1-\alpha = \sum_{Q_0 \supset s} p(s),$$

donde Q_0 es el conjunto de todas las posible muestras cuyo intervalo de confianza contiene al total poblacional t_y. En la práctica muy pocas veces se conoce la varianza del estimador; por lo tanto, el intervalo de confianza estimado de nivel $(1-\alpha)$ puede ser obtenido con los datos de la muestra seleccionada reemplazando en (2.2.8) la varianza del estimador por su correspondiente estimación y tomaría la siguiente expresión

$$IC_s(1-\alpha) = \left[\hat{t}_{y,\pi} - z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{t}_{y,\pi})}, \hat{t}_{y,\pi} + z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{t}_{y,\pi})}\right]. \qquad (2.2.9)$$

Al utilizar una estrategia de muestreo en la estimación de un parámetro en poblaciones finitas, las propiedades de la estrategia se estudian en términos de:

- **Confiabilidad:** definida como la suma de las probabilidades de las muestras cuyo intervalo de confianza contiene al parámetro.
- **Precisión:** definida como la longitud del intervalo de confianza.

Nótese que las anteriores propiedades están en función del intervalo de confianza. Para determinar la confiabilidad se debe conocer al parámetro T (desconocido) por tanto, en términos prácticos la confiabilidad no se puede calcular. Para determinar la precisión y la confiabilidad se requiere conocer la varianza, basada en el diseño de muestreo, del estimador utilizado, digamos $\hat{T}$; sin embargo, el cálculo de esta varianza $Var(\hat{T})$ implica, casi siempre, el requerimiento de conocer los valores y_k para todo $k = 1, ..., N$. Luego la precisión tampoco se puede calcular. Sin embargo se debe proponer un estimador de $Var(\hat{T})$ (ojalá insesgado) que junto con $\hat{T}$ proporción una cota para el sesgo y para la precisión.

Estimación de otros parámetros

Aunque (2.2.2) es un estimador del total poblacional de la característica de interés, se puede utilizar para estimar otras cantidades poblacionales de interés. Si

el tamaño poblacional N es conocido, la media poblacional definida en (2.1.14) puede ser estimada con el estimador de Horvitz-Thompson.

Resultado 2.2.6. *La media poblacional es estimada insesgadamente mediante el uso de la siguiente expresión*

$$\hat{\bar{y}}_\pi = \frac{1}{N}\left(\hat{t}_{y,\pi}\right) = \frac{1}{N}\sum_s \frac{y_k}{\pi_k} \tag{2.2.10}$$

La varianza y la varianza estimada del estimador de la media poblacional están dadas por

$$Var(\hat{\bar{y}}_\pi) = \frac{1}{N^2}Var(\hat{t}_{y,\pi}) \tag{2.2.11}$$

$$\hat{Var}(\hat{\bar{y}}_\pi) = \frac{1}{N^2}\hat{Var}(\hat{t}_{y,\pi}) \tag{2.2.12}$$

respectivamente,

Sin embargo, es la regla más que la excepción que en la mayoría de casos en donde el usuario se enfrenta a una investigación cuyos objetivos están supeditados a la realización de un estudio por muestreo que el tamaño poblacional sea desconocido. En tal caso, podemos usar el estimador de Horvitz-Thompson para estimarlo puesto que N puede ser escrito de la siguiente manera

$$N = \sum_U 1, \tag{2.2.13}$$

tomando la conocida forma de un total poblacional. Luego, tenemos el siguiente resultado.

Resultado 2.2.7. *El tamaño poblacional es estimado insesgadamente mediante el uso de la siguiente expresión*

$$\hat{N}_\pi = \sum_S \frac{1}{\pi_k}. \tag{2.2.14}$$

Cuando se ha estimado el total poblacional de una característica de interés y el tamaño poblacional mediante el uso del estimador de Horvitz-Thompson, surge un estimador para la media poblacional dado por

$$\widetilde{y}_S = \frac{\hat{t}_{y,\pi}}{\hat{N}_\pi} \tag{2.2.15}$$

$$= \sum_S \frac{y_k}{\pi_k} \Big/ \sum_S \frac{1}{\pi_k}. \tag{2.2.16}$$

La anterior expresión es una razón, o un cociente entre dos totales poblacionales. Las propiedades estadísticas de los anteriores estimadores serán tratados más adelante en las secciones pertinentes del libro.

Tillé (2006) cita que aun al conocer N, una mala propiedad del estimador de Horvitz-Thompson para la media poblacional se tiene al utilizarlo cuando la característica de interés es constante para todos los elementos de la población ($y_k = C$

$\forall k \in U$). Por supuesto, bajo las anteriores condiciones es claro que la media poblacional es igual a la constante ($\bar{y}_U = C$). Sin embargo, el estimador $\hat{\bar{y}}_\pi$ toma la siguiente forma

$$\hat{\bar{y}}_\pi = \frac{1}{N}\sum_s \frac{y_k}{\pi_k} = \frac{1}{N}\sum_s \frac{C}{\pi_k} = \frac{C}{N}\sum_s \frac{1}{\pi_k} = C\frac{\hat{N}_\pi}{N}. \tag{2.2.17}$$

Al respecto, Bautista (1998) afirma que en aquellos casos en los que se conoce el valor de N es preferible ignorarlo y utilizar el estimador $\widetilde{y}_S$ puesto que su variación es menor y cuando $y_k = C \ \forall k \in U$ reproduce la media poblacional con varianza nula puesto que

$$\widetilde{y}_S = \frac{\hat{t}_{y,\pi}}{\hat{\bar{y}}_\pi} = \frac{C\hat{\bar{y}}_\pi}{\hat{\bar{y}}_\pi} = C.$$

Cuando el tamaño poblacional es conocido y, como se verá más adelante, para algunos diseños de muestreo sin reemplazo, se puede crear un nuevo estimador alternativo del total poblacional inspirado en el siguiente argumento: Si $\widetilde{y}_S$ estima la media poblacional, entonces $N\widetilde{y}_S$ estimará el total poblacional. Por tanto, el estimador alternativo está dado por la siguiente expresión

$$\hat{t}_{y,alt} = N\widetilde{y}_S = \hat{t}_{y,\pi}\frac{N}{\hat{N}_\pi} \tag{2.2.18}$$

que se puede ver como una corrección del estimador de Horvitz-Thompson mediante la estimación del tamaño de la población. La varianza y la estimación de la varianza serán tema de capítulos posteriores.

Ejemplo 2.2.1. La función `HT` del paquete `TeachingSampling` arroja la estimación del total poblacional de una o varias características de interés. Esta función tiene dos argumentos: el vector de tamaño n de probabilidades de inclusión `pik` y el conjunto de valores de la característica o características de interés en los individuos pertenecientes a la muestra, `y` puede ser un vector en el caso de una sola característica de interés o una matriz en el caso de varias.

Así, si la primera muestra (cuyos elementos son **Yves** y **Ken**) hubiese sido seleccionada y dado que las probabilidades de inclusión de estos dos elementos son 0.58 y 0.34, respectivamente y los valores de la característica de interés son 32 y 34, respectivamente, el estimador de Horvitz-Thompson arrojaría la siguiente estimación:

```
> ys <- c(32,34)
> pik <- c(0.58,0.34)
> HT(ys,pik)
         [,1]
[1,] 155.1724
```

Nótese que el total poblacional para la variable de interés y es igual a 236. Por otro lado, el cálculo o estimación de la varianza del estimador de Horvitz-Thompson no se encuentra implementado pues la doble suma hace que los procesos computacionales sean muy largos y demorado. Por tanto, si se quieren conocer estos valores,

el proceso se debe realizar manualmente. La estimación de la varianza se realiza teniendo en cuenta que $\pi_{12} = 0.13$. Así,

$$\frac{\Delta_{11}}{\pi_{11}} = \frac{\pi_{11} - \pi_1\pi_1}{\pi_{11}} = \frac{0.58 - 0.58^2}{0.58} = 0.42$$
$$\frac{\Delta_{12}}{\pi_{12}} = \frac{\pi_{12} - \pi_1\pi_2}{\pi_{12}} = \frac{0.13 - 0.58 * 0.34}{0.13} = -0.52$$
$$\frac{\Delta_{21}}{\pi_{21}} = \frac{\pi_{11} - \pi_2\pi_1}{\pi_{21}} = \frac{0.13 - 0.34 * 0.58}{0.13} = -0.52$$
$$\frac{\Delta_{22}}{\pi_{22}} = \frac{\pi_{22} - \pi_2\pi_2}{\pi_{22}} = \frac{0.34 - 0.34^2}{0.34} = 0.66$$

Por tanto, utilizando (2.2.6), el estimador de la varianza será

$$\widehat{Var}(\hat{t}_\pi) = \frac{\Delta_{11}}{\pi_{11}}\frac{y_1}{\pi_1}\frac{y_1}{\pi_1} + \frac{\Delta_{12}}{\pi_{12}}\frac{y_1}{\pi_1}\frac{y_2}{\pi_2} + \frac{\Delta_{21}}{\pi_{21}}\frac{y_2}{\pi_2}\frac{y_1}{\pi_1} + \frac{\Delta_{22}}{\pi_{22}}\frac{y_2}{\pi_2}\frac{y_2}{\pi_2}$$

y su respectiva estimación será

$$0.42\left(\frac{32}{0.58}\right)^2 - 2(0.52)\left(\frac{32}{0.58}\frac{34}{0.34}\right) + 0.66\left(\frac{34}{0.34}\right)^2 \cong 2140$$

El coeficiente de variación estimado es

$$cve(\hat{t}_\pi) = \frac{\sqrt{2140}}{155.1724} \cong 0.3$$

Y el intervalo de confianza estimado con un nivel de confianza del 95 por ciento para esta estimación es el siguiente:

$$\begin{aligned} IC_s(0.95) &\cong \left[155 - (1.96)\sqrt{2140}, 155 + (1.96)\sqrt{2140}\right] \\ &\cong [64, 246] \end{aligned}$$

Continuando con el ejercicio léxico-gráfico de la estimación del total poblacional t_y en todas las posibles muestras de tamaño 10 de la población U, tenemos la siguiente tabla.

	X1	X2	p	pi.k	pi.l	y.k	y.l	Est
1	Yves	Ken	0.13	0.58	0.34	32	34	155.2
2	Yves	Erik	0.20	0.58	0.48	32	46	151.0
3	Yves	Sharon	0.15	0.58	0.33	32	89	324.8
4	Yves	Leslie	0.10	0.58	0.27	32	35	184.8
5	Ken	Erik	0.15	0.34	0.48	34	46	195.8
6	Ken	Sharon	0.04	0.34	0.33	34	89	369.6

```
7      Ken Leslie     0.02     0.34 0.27       34  35     229.6
8     Erik Sharon     0.06     0.48 0.33       46  89     365.5
9     Erik Leslie     0.07     0.48 0.27       46  35     225.4
10 Sharon Leslie     0.08     0.33 0.27       89  35     399.3
```

El vector Est contiene las estimaciones Horvitz-Thompson para cada una de las 10 posibles muestras, su esperanza se calcula como

```
> sum(p*est)
[1] 236
```

Nótese que la esperanza del estimador de Horvitz-Thompson reproduce exactamente el total poblacional. La varianza se calcula de la siguiente manera

$$Var(\hat{t}_\pi) = (0.13)(155.2-236)^2 + (0.2)(151.0-236)^2 + \cdots \\ + (0.08)(399.3-236)^2 = 7847.2$$

Acudiendo a la función VarHT, del paquete TeachignSampling, es posible reproducir este mismo calculo de la varianza. Sin embargo, esta función utiliza la expresión teórica de la varianza $Var_1(\hat{t}_{y,\pi})$ dada por (2.2.4) para diseños de muestreo de tamaño fijo. Tiene cuatro argumentos: y, que es un vector que contiene los valores de la característica de interés en todos y cada uno de los elementos de la población; N, el tamaño de la población; n, el tamaño de muestra fijo y p, el diseño de muestreo utilizado. El resultado de esta función es el cálculo del valor de la varianza teórica del estimador de Horvitz-Thompsosn para un diseño de muestreo y una configuración de valores poblacionales particular. Siguiendo con el diseño de muestreo dado en el ejemplo 2.1.2 y la configuración de valores de la característica de interés del ejemplo 2.1.3, tenemos que el calculo de la varianza es exactamente igual al dado por el ejercicio léxico-gráfico.

```
> VarHT(y1, N, n, p)
[1] 7847.2
```

2.2.2 El estimador de Hansen-Hurwitz

Sobre el muestreo con reemplazo

Considere una población finita de N elementos y un diseño de muestreo que permite la selección de una muestra realizada s, con reemplazo, de tamaño m. Como Lohr (2000) lo afirma, la manera más intuitiva de entender este tipo de diseños muestrales con reemplazo es pensar en la extracción de m muestras independientes de tamaño 1. Se extrae un elemento de la población para ser incluido en la muestra con una probabilidad p_k; sin embargo, ese mismo elemento participa en el siguiente sorteo aleatorio. Este proceso se repite m veces; es decir, se tiene un total de m sorteos aleatorios.

Bajo el anterior esquema de selección, es claro que un elemento puede ser seleccionado en la muestra más de una vez; por lo tanto, aunque el tamaño de la muestra

seleccionada con reemplazo es m, el tamaño de muestra efectivo no es necesariamente m. Nótese que la selección de un elemento que se repite más de una vez no proporciona información nueva. Es por esto que en la práctica, se prefieren los diseños de muestreo que permita la selección de muestras sin duplicados.

Särndal, Swensson & Wretman (1992) afirman que el marco general del muestreo con reemplazo tiene las siguientes características:

- Cada elemento de la población está relacionado directamente con un número positivo p_k $(k = 1, \ldots, N)$ de tal forma que
$$\sum_U p_k = 1.$$
A p_k se le conoce como la **probabilidad de selección** del elemento k-ésimo. Nótese que estas probabilidades no son necesariamente iguales.
- Para seleccionar el primer elemento que pertenecerá a la muestra de tamaño m, se lleva a cabo un sorteo aleatorio de tal forma que
$$Pr(\text{Seleccionar el elemento } k) = p_k, \ k \in U.$$
- El elemento seleccionado es reemplazado en la población y vuelva a ser parte del próximo sorteo aleatorio con la misma probabilidad de selección p_k.
- El mismo conjunto de probabilidades es usado para seleccionar los restantes elementos. En total se realizan m sorteos aleatorios independientes.

Ahora, en muestreo con reemplazo la probabilidad de selección de un elemento no es lo mismo que la probabilidad de inclusión[3] del mismo. Se tienen los siguientes resultados.

Definición 2.2.1. *Bajo un diseño con reemplazo, se define la variable aleatoria $n_k(S)$ como el número de veces que el elemento k-ésimo es seleccionado en la muestra aleatoria S.*

Resultado 2.2.8. *La variable aleatoria $n_k(S)$ sigue una distribución binomial tal que*
$$E(n_k(S)) = mp_k, \qquad Var(n_k(S)) = mp_k(1 - p_k)$$

Prueba. Dado que cada una de las m extracciones inducen eventos estadísticos independientes, la selección en una extracción particular del k-ésimo elemento sigue una distribución de Bernoulli, con parámetro p_k. Como se trata de m extracciones, $n_k(S)$ sigue una distribución binomial y puede tomar los valores $0, 1, \ldots, m$; al definir éxito como la selección del elemento k-ésimo en la muestra, entonces se tiene la demostración del resultado. ■

Definición 2.2.2. *De manera general, un diseño de muestreo con reemplazo se define como*
$$p(s) = \begin{cases} \frac{m!}{n_1(s)!\ldots n_N(s)!} \prod_U (p_k)^{n_k(s)} & \textit{si } \sum_U n_k(s) = m \\ 0 & \textit{en otro caso} \end{cases} \tag{2.2.19}$$

[3]Nótese que la probabilidad de inclusión se refiere a la probabilidad de que el elemento sea seleccionado al menos una vez en la muestra.

Donde $n_k(s)$ es el número de veces que el elemento k-ésimo es seleccionado en la muestra realizada s.

Nótese la diferencia (y a la vez similitud) de la variable $n_k(S)$ con la variable $I_k(S)$, además por la definición anterior se tiene que el diseño de muestreo con reemplazo sigue una distribución multinomial, por lo tanto cumple las condiciones de diseño muestral; es decir, $\sum_{s \in Q} p(s) = 1$, donde Q es el soporte que contiene todas las posibles muestras con reemplazo de tamaño m. La cardinalidad de Q, es

$$\#Q = \binom{N+m-1}{m} \tag{2.2.20}$$

Resultado 2.2.9. *En muestreo con reemplazo, la probabilidad de inclusión de primer orden del elemento k-ésimo está dada por:*

$$\pi_k = 1 - (1 - p_k)^m \tag{2.2.21}$$

Prueba. Dado que se trata de eventos independientes los cuales tienen asociada una probabilidad de éxito (éxito equivalente a que el elemento $k \in s)p_k$, entonces cada uno de estos sorteos aleatorios está determinado por una distribución de probabilidad de tipo Bernoulli. Por consiguiente, cuando se realizan m ensayos independientes, se utiliza la distribución de probabilidad binomial para hallar las probabilidades de inclusión de primer orden de cada uno de los elementos en la población

$$\begin{aligned}
\pi_k = Pr(k \in S) &= 1 - Pr(k \notin s)\\
&= 1 - \binom{m}{m}(1-p_k)^m (p_k)^{m-m}\\
&= 1 - (1-p_k)^m
\end{aligned}$$

■

Resultado 2.2.10. *En muestreo con reemplazo, las probabilidades de inclusión de segundo orden π_{kl}, están dadas por:*

$$\pi_{kl} = 1 - (1-p_k)^m - (1-p_l)^m + (1 - p_k - p_l)^m \qquad k \neq l = 1 \ldots, N \tag{2.2.22}$$

Prueba. Para hallar esta probabilidad debemos negar que ($k \in S$ y $l \in s$). Esta negación da como resultado ($k \notin s$ ó $l \notin s$). Suponga que tenemos dos eventos, $A = (k \notin s)$ y $B = (l \notin s)$; por tanto, $Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B)$. Las probabilidades anteriores se rigen por un modelo binomial, luego:

$$\begin{aligned}
\pi_{kl} &= Pr(k \in S \text{ y } l \in s)\\
&= 1 - Pr(k \notin s) - Pr(l \notin s) + Pr(k, l \notin s)\\
&= 1 - (1-p_k)^m - (1-p_l)^m + \binom{m}{m}(1-p_k-p_l)^m (p_k + p_l)^{m-m}\\
&= 1 - (1-p_k)^m - (1-p_l)^m + (1-p_k-p_l)^m
\end{aligned}$$

El cuarto sumando en la igualdad anterior se obtiene considerando que cada ensayo se toma como un proceso Bernoulli, donde el éxito es *no escoger ni a k ni a l*. Por tanto

$$
\begin{aligned}
Pr(\text{Éxito}) &= 1 - Pr(\text{Fracaso}) \\
&= 1 - Pr(\text{Escoger a } k) - Pr(\text{Escoger a } l) + Pr(\text{Escoger a ambos}) \\
&= 1 - p_k - pl
\end{aligned}
$$

Puesto que se trata de un sólo ensayo, la probabilidad de escoger a ambos es nula. ■

Esto se nota más claramente con el típico ejemplo del dado. Si el evento es el lanzamiento de un dado y el éxito es *no sacar 3 o 5*, entonces la probabilidad de obtener éxito será: $1 - Pr(\text{Fracaso})$, es decir $1 - Pr(\text{Sale } 5) - Pr(\text{Sale } 1) + Pr(\text{Sale } 5 \text{ y } 1)$. Es obvio que el último sumando es cero dado que se trata de un sólo lanzamiento.

Ejemplo 2.2.2. El lector no debe confundir el concepto de **muestra con reemplazo** con el concepto de **extracción ordenada**. En nuestra población ejemplo el tamaño poblacional es $N = 5$. Si se utiliza un diseño de muestreo que induzca muestras de tamaño fijo igual a $m = 2$, entonces existirían $N^m = 5^2 = 25$ posibles extracciones ordenadas. Sin embargo, sólo existen $\binom{N+m-1}{m} = \binom{6}{2} = 15$ posibles muestras con reemplazo. Este escenario es evidenciado fácilmente con la ayuda de la variable aleatoria $n_k(S)$. Las posibles extracciones ordenadas están dadas de la siguiente manera.

```
(1,1)   (2,1)   (3,1)   (4,1)   (5,1)
(1,2)   (2,2)   (3,2)   (4,2)   (5,2)
(1,3)   (2,3)   (3,3)   (4,3)   (5,3)
(1,4)   (2,4)   (3,4)   (4,4)   (5,4)
(1,5)   (2,5)   (3,5)   (4,5)   (5,5)
```

Sin embargo, aunque todas las posibles extracciones ordenas no constituyen el soporte de muestreo, éstas si ayudan a definirlo. De hecho, el primer paso para la construcción del soporte de muestreo con reemplazo es la determinación de todas las posibles extracciones. La función `OrderWR`[4] del paquete `TeachingSampling` permite conocer todas las posibles extracciones de tamaño fijo para un diseño de muestreo con reemplazo.

Esta función cuenta con tres argumentos: el primer argumento correspondiente al tamaño de la población `N`, el segundo, correspondiente al tamaño de las selecciones, `m`, que no necesariamente debe ser menor que el tamaño poblacional[5] y, el último corresponde a una característica `ID` que puede ser un conjunto de rótulos o cualquier otro tipo de identificador continuo. El resultado de la función `OrderWR` será un conjunto de todas las posibles extracciones ordenadas con tamaño fijo `m`.

[4]El autor desea recalcar que el resultado de esta función no corresponde al soporte de muestreo con reemplazo sino al conjunto de todas las posibles extracciones ordenadas con reemplazo y de tamaño fijo.

[5]Se enfatiza que para este tipo de diseños de muestreo con reemplazo es posible que el tamaño de muestra sea mayor al tamaño poblacional.

Cuando el argumento `ID` es distinto de `FALSE`, la salida de la función corresponderá al rótulo o identificador continuo para cada elemento de la población. En el siguiente ejemplo se utiliza esta función en nuestra población ejemplo U.

```
> N <- length(U)
> N
[1] 5
> m <- 2
```

```
>OrderWR(N,m,ID=FALSE)
      [,1] [,2]
 [1,]    1    1
 [2,]    1    2
 [3,]    1    3
 [4,]    1    4
 [5,]    1    5
 [6,]    2    1
 [7,]    2    2
 [8,]    2    3
 [9,]    2    4
[10,]    2    5
[11,]    3    1
[12,]    3    2
[13,]    3    3
[14,]    3    4
[15,]    3    5
[16,]    4    1
[17,]    4    2
[18,]    4    3
[19,]    4    4
[20,]    4    5
[21,]    5    1
[22,]    5    2
[23,]    5    3
[24,]    5    4
[25,]    5    5
```

```
> OrderWR(N,m,ID=U)
      [,1]     [,2]
 [1,] "Yves"   "Yves"
 [2,] "Yves"   "Ken"
 [3,] "Yves"   "Erik"
 [4,] "Yves"   "Sharon"
 [5,] "Yves"   "Leslie"
 [6,] "Ken"    "Yves"
 [7,] "Ken"    "Ken"
 [8,] "Ken"    "Erik"
 [9,] "Ken"    "Sharon"
[10,] "Ken"    "Leslie"
[11,] "Erik"   "Yves"
[12,] "Erik"   "Ken"
[13,] "Erik"   "Erik"
[14,] "Erik"   "Sharon"
[15,] "Erik"   "Leslie"
[16,] "Sharon" "Yves"
[17,] "Sharon" "Ken"
[18,] "Sharon" "Erik"
[19,] "Sharon" "Sharon"
[20,] "Sharon" "Leslie"
[21,] "Leslie" "Yves"
[22,] "Leslie" "Ken"
[23,] "Leslie" "Erik"
[24,] "Leslie" "Sharon"
[25,] "Leslie" "Leslie"
```

Nótese que el conjunto de extracciones ordenadas contiene al soporte de muestreo con reemplazo. Sin embargo, con ayuda de la función `SupportWR` del paquete `TeachingSampling` se define el verdadero soporte inducido por el diseño de muestreo con reemplazo. Los argumentos de esta función son los mismos tres de la función `OrderWR`: `N`, `m` y `ID`. El resultado de la función es el conjunto de todas las posibles muestras con reemplazo de tamaño fijo. Para este ejemplo particular, el soporte está dado por las siguientes muestras y no por todas las posibles extracciones ordenadas.

```
> SupportWR(N, m,ID=FALSE)

      [,1] [,2]
 [1,]    1    1
 [2,]    1    2
 [3,]    1    3
 [4,]    1    4
 [5,]    1    5
 [6,]    2    2
 [7,]    2    3
 [8,]    2    4
 [9,]    2    5
[10,]    3    3
[11,]    3    4
[12,]    3    5
[13,]    4    4
[14,]    4    5
[15,]    5    5
```

```
> SupportWR(N,m,ID=U)

      [,1]     [,2]
 [1,] "Yves"   "Yves"
 [2,] "Yves"   "Ken"
 [3,] "Yves"   "Erik"
 [4,] "Yves"   "Sharon"
 [5,] "Yves"   "Leslie"
 [6,] "Ken"    "Ken"
 [7,] "Ken"    "Erik"
 [8,] "Ken"    "Sharon"
 [9,] "Ken"    "Leslie"
[10,] "Erik"   "Erik"
[11,] "Erik"   "Sharon"
[12,] "Erik"   "Leslie"
[13,] "Sharon" "Sharon"
[14,] "Sharon" "Leslie"
[15,] "Leslie" "Leslie"
```

Por supuesto, cada una de las posibles muestras con reemplazo que pertenecen al soporte tiene distintas probabilidades de selección dependiendo de la configuración de las probabilidades de selección individuales para cada elemento, p_k. Supongamos que cada uno de los cinco elementos de la población tiene probabilidad de selección dadas por

$$p_k = \begin{cases} 1/4, & \text{para } k = \textbf{Yves, Ken, Leslie}, \\ 1/8, & \text{para } k = \textbf{Sharon, Erik} \end{cases}$$

Nótese que $\sum_U p_k = 1$. Para esta configuración particular, y siguiendo la expresión (2.2.19), las probabilidades de selección $p(s)$ de las muestras en el soporte y el valor de la variable $n_k(S)$ estarían dadas por la siguiente configuración.

```
     V1       V2        p(s)        n1 n2 n3 n4 n5
1    Yves     Yves      0.0625      2  0  0  0  0
2    Yves     Ken       0.125       1  1  0  0  0
3    Yves     Erik      0.0625      1  0  1  0  0
4    Yves     Sharon    0.0625      1  0  0  1  0
5    Yves     Leslie    0.125       1  0  0  0  1
6    Ken      Ken       0.0625      0  2  0  0  0
7    Ken      Erik      0.0625      0  1  1  0  0
8    Ken      Sharon    0.0625      0  1  0  1  0
9    Ken      Leslie    0.125       0  1  0  0  1
10   Erik     Erik      0.015625    0  0  2  0  0
11   Erik     Sharon    0.03125     0  0  1  1  0
12   Erik     Leslie    0.0625      0  0  1  0  1
13   Sharon   Sharon    0.015625    0  0  0  2  0
14   Sharon   Leslie    0.0625      0  0  0  1  1
15   Leslie   Leslie    0.0625      0  0  0  0  2
```

Nótese que la suma de las probabilidades de selección inducidas por el diseño de muestreo es igual a uno y que cada una de ellas es mayor que cero. El lector debe fijarse en que la muestra perteneciente al soporte está dada en términos de $n_k(S)$. De esta manera, si se ha seleccionado la séptima muestra dada por `1 0 1 0 0`, en realidad, no importa si **Yves** fue seleccionado primero o después que **Erik** y la probabilidad de selección de esta muestra particular es 0.125 pues

$$\begin{aligned} p(s) &= \frac{2!}{1!0!1!0!0!}\left[\left(\frac{1}{4}\right)^1\left(\frac{1}{4}\right)^0\left(\frac{1}{8}\right)^1\left(\frac{1}{8}\right)^0\left(\frac{1}{4}\right)^0\right] \\ &= 2\left(\frac{1}{32}\right) = 0.0625 \end{aligned}$$

Estimador del total poblacional

Hansen, Hurwitz & Madow (1953) proponen un estimador conveniente para el total de una población t_y cuando el diseño de muestreo es con reemplazo. La lógica que sigue en la construcción de este estimador está dada a continuación. Sea el evento aleatorio:

Seleccionar el elemento k ($k \in U$) en el i-ésimo sorteo ($i = 1, \ldots, m$).

Este evento define la creación de variables aleatorias, que serán utilizadas más adelante, cuyo comportamiento es posible modelar mediante el siguiente resultado.

Resultado 2.2.11. *Sean $U_1, U_2, \ldots, U_m$ es una sucesión de variables aleatorias independientes e idénticamente distribuidas con $E(U_i) = \mu$ y $Var(U_i) = \sigma^2$. Sea $\bar{U} = \sum_{i=1}^m U_i/m$. Entonces $E(\bar{U}) = \mu$, $Var(\bar{U}) = \sigma^2/m$ y un estimador insesgado de $Var(\bar{U})$ está dado por la siguiente expresión*

$$\widehat{Var}(\bar{U}) = \frac{1}{m(m-1)}\sum_{i=1}^m (U_i - \bar{U})^2 \tag{2.2.23}$$

y por consiguiente, un estimador insesgado para σ^2 está dado por

$$\hat{\sigma^2} = \frac{1}{m-1}\sum_{i=1}^m (U_i - \bar{U})^2. \tag{2.2.24}$$

Prueba. La esperanza de $\bar{U}$ es

$$E(\bar{U}) = \frac{1}{m}\sum_{i=1}^m E(U_i) = \mu \tag{2.2.25}$$

La varianza está determinada por

$$Var(\bar{U}) = \frac{1}{m^2}\sum_{i=1}^m Var(U_i) = \sigma^2/m \tag{2.2.26}$$

Nótese que los términos de covarianza son nulos puesto que las variables son independientes entre ellas. Ahora como

$$\sum_{i=1}^{m}(U_i - \bar{U})^2 = \sum_{i=1}^{m} U_i^2 - m\bar{U}^2 \tag{2.2.27}$$

entonces,

$$E(\sum_{i=1}^{m}(U_i - \bar{U})^2) = \sum_{i=1}^{m} E(U_i^2) - mE(\bar{U}^2) \tag{2.2.28}$$

Por otro lado

$$E(U_i^2) = Var(U_i) + [E(U_i)]^2 = \sigma^2 + \mu^2$$
$$E(\bar{U}^2) = Var(\bar{U}) + [E(\bar{U})]^2 = \sigma^2 / m + \mu^2$$

Esto conduce a la demostración del teorema puesto que

$$E(\sum_{i=1}^{m}(U_i - \bar{U})^2) = (m-1)\sigma^2 \tag{2.2.29}$$

■

El anterior es un resultado muy potente que puede ser utilizado para cualquier tipo de variables aleatorias que sean independientes e idénticamente distribuidas y será la base para la demostración de resultados en la estimación de parámetros que utilicen diseños de muestreo con reemplazo. Siguiendo con el marco teórico del muestreo con reemplazo tenemos la siguiente definición.

Definición 2.2.3. *Se define la variable aleatoria Z_i tal que*

$$Z_i = y_{k_i}/p_{k_i} \qquad k \in U \quad i = 1, \ldots, m \tag{2.2.30}$$

donde la cantidad y_{k_i} es el valor de la característica de interés del k-ésimo elemento seleccionado en la i-ésima extracción. Análogamente, p_{k_i} es el valor de la probabilidad de selección del k-ésimo elemento seleccionado en la i-ésima extracción.

Resultado 2.2.12. *La distribución de la variable aleatoria Z_i está dada por*

$$Pr\left(Z_i = \frac{y_k}{p_k}\right) = p_k, \tag{2.2.31}$$

por tanto la esperanza y varianza de la variable aleatoria Z_i son

$$E(Z_i) = t_y \tag{2.2.32}$$

y

$$Var(Z_i) = \sum_U p_k \left(\frac{y_k}{p_k} - t_y\right)^2, \tag{2.2.33}$$

respectivamente.

Prueba. Dado que se trata de m sorteos aleatorios independientes, la variable aleatoria Z_i puede tomar los siguientes valores

$$\frac{y_1}{p_1}, \frac{y_2}{p_2} \ldots, \frac{y_N}{p_N}$$

con probabilidades

$$p_1, p_2 \ldots, p_N$$

respectivamente. Luego, acudiendo a la definición genérica del operador esperanza, se tiene

$$E(Z_i) = \sum_U \frac{y_k}{p_k} Pr\left(Z_i = \frac{y_k}{p_k}\right) = \sum_U \frac{y_k}{p_k} p_k = t_y$$

y análogamente se tiene la varianza

$$Var(Z_i) = \sum_U \left(\frac{y_k}{p_k} - E(Z_i)\right)^2 Pr\left(Z_i = \frac{y_k}{p_k}\right) = \sum_U \left(\frac{y_k}{p_k} - t_y\right)^2 p_k$$

■

Dado que las m extracciones son eventos independientes, también lo son las variables Z_i[6]. Nótese que la cantidad Z_i es una estimación del total poblacional con la i-ésima muestra seleccionada de tamaño 1. Ahora, como existen m sorteos habrán m estimaciones del total poblacional; por tanto, como en mucho otros procedimientos estadísticos utilizamos el promedio de estas m estimaciones para obtener una estimación unificada para t_y. El estimador de Hansen-Hurwitz toma la siguiente forma

$$\hat{t}_{y,p} = \frac{1}{m} \sum_{i=1}^{m} \frac{y_{k_i}}{p_{k_i}} \tag{2.2.34}$$

Para tener una estrategia de muestreo que resulte eficiente en la estimación de t_y, es conveniente utilizar el estimador de Hansen-Hurwitz, cuando las probabilidades de selección son proporcionales a la característica de interés; esto es, cuando $p_k \propto y_k$. Si lo anterior sucede, el estimador tendrá una varianza casi nula y la estimación será muy precisa.

Resultado 2.2.13. *Si $p_k > 0$, para todo $k \in U$, el estimador $\hat{t}_{y,p}$ es insesgado*

Prueba. Las variables aleatorias Z_i son independientes (porque cada ensayo es independiente) y su distribución está inducida por $Pr(Z_i = y_k/p_k) = p_k$, $k \in U$; es decir, son idénticamente distribuidas. Por tanto, el estimador de Hansen-Hurwitz puede escribirse como:

$$\hat{t}_{y,p} = \frac{1}{m} \sum_{i=1}^{m} \frac{y_i}{p_i} = \frac{1}{m} \sum_{i=1}^{m} Z_i = \bar{Z}$$

[6] $Z_1, \ldots, Z_m$ define una sucesión de variables aleatorias independientes e idénticamente distribuidas, o si se quiere, en términos de la inferencia clásica, define una **muestra aleatoria**.

y así con $p_k > 0$ para todo $k \in U$, tenemos

$$E(\hat{t}_{y,p}) = \frac{1}{m}\sum_{i=1}^{m} E(Z_i) = \frac{1}{m}\sum_{i=1}^{m} t_y = t_y$$

■

Varianza del estimador de Hansen-Hurwitz

Una de las características más importantes del estimador de Hansen-Hurwitz es la sencillez de la expresión de su varianza. Esta misma hace que aunque el muestreo sea con reemplazo, el estimador de Hansen-Hurwitz sea utilizado de manera frecuente por los usuarios de los estudios por muestreo.

Resultado 2.2.14. *La varianza del estimador de Hansen-Hurwitz está dada por la siguiente expresión*

$$Var(\hat{t}_{y,p}) = \frac{1}{m}\sum_{k=1}^{N} p_k \left(\frac{y_k}{p_k} - t_y\right)^2 \tag{2.2.35}$$

Prueba. Por la independencia de las selecciones se tiene que

$$\begin{aligned} Var(\hat{t}_{y,p}) &= Var\left(\frac{1}{m}\sum_{i=1}^{m} Z_i\right) \\ &= \frac{1}{m^2}\sum_{i=1}^{m} Var(Z_i) \\ &= \frac{1}{m} Var(Z_i) \\ &= \frac{1}{m}\sum_{U}\left(\frac{y_k}{p_k} - t_y\right)^2 p_k \end{aligned}$$

■

La anterior expresión hace que el cálculo computacional de la varianza del estimador de Hansen-Hurwitz sea muy sencillo. Sin embargo, esta varianza se puede escribir de varias formas, algunas de ellas muy útiles para el desarrollo teórico de las propiedades del estimador.

Resultado 2.2.15. *De manera general, la varianza del estimador de Hansen-Hurwitz se puede escribir de la siguiente manera*

$$Var(\hat{t}_{y,p}) = \frac{1}{m}\left(\sum_{k=1}^{N} \frac{y_k^2}{p_k} - t_y^2\right) \tag{2.2.36}$$

Prueba.

$$
\begin{aligned}
Var(\hat{t}_{y,p}) = &= \frac{1}{m}\sum_{k=1}^{N} p_k\left(\frac{y_k}{p_k} - t_y\right)^2 \\
&= \frac{1}{m}\sum_{k=1}^{N} p_k\left(\frac{y_k^2}{p_k^2} - 2t_y\frac{y_k}{p_k} + t_y^2\right) \\
&= \frac{1}{m}\sum_{k=1}^{N}\left(\frac{y_k^2}{p_k} - 2t_y y_k + p_k t_y^2\right) \\
&= \frac{1}{m}\left(\sum_{k=1}^{N}\frac{y_k^2}{p_k} - 2t_y\sum_{k=1}^{N} y_k + t_y^2\sum_{k=1}^{N} p_k\right) \\
&= \frac{1}{m}\left(\sum_{k=1}^{N}\frac{y_k^2}{p_k} - 2t_y^2 + t_y^2\right) = \frac{1}{m}\left(\sum_{k=1}^{N}\frac{y_k^2}{p_k} - t_y^2\right)
\end{aligned}
$$

■

Estimación de la varianza

Resultado 2.2.16. *Un estimador insesgado de la expresión (2.2.35) es*

$$
\widehat{Var}(\hat{t}_{y,p}) = \frac{1}{m(m-1)}\sum_{i=1}^{m}\left(\frac{y_i}{p_i} - \hat{t}_{y,p}\right)^2 \tag{2.2.37}
$$

Prueba. Al desarrollar la varianza del estimador llegamos a que ésta es igual a

$$
\frac{1}{m}Var(Z_i).
$$

Ahora, utilizando el resultado 2.2.11, como $Z_1, \ldots, Z_m$ conforman una muestra aleatoria de variables con esperanza t_y e idéntica varianza, entonces un estimador natural e insesgado para la varianza de Z_i es

$$
\frac{1}{m-1}\sum_{i=1}^{m}(Z_i - \bar{Z})^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(\frac{y_i}{p_i} - \hat{t}_{y,p}\right)^2
$$

por tanto, un estimador insesgado de la varianza del estimador de Hansen-Hurwitz será

$$
\widehat{Var}(\hat{t}_{y,p}) = \frac{1}{m}\frac{1}{m-1}\sum_{i=1}^{m}\left(\frac{y_{k_i}}{p_{k_i}} - \hat{t}_{y,p}\right)^2
$$

■

Resultado 2.2.17. *Una expresión alternativa para la estimación de la varianza del estimador de Hansen-Hurwitz en muestreo con reemplazo es*

$$
\widehat{Var}(\hat{t}_{y,p}) = \frac{1}{m(m-1)}\sum_{i=1}^{m}\left(\frac{y_{k_i}}{p_{k_i}}\right)^2 - m\hat{t}_{y,p}^2
$$

Prueba. Partiendo del resultado anterior, se tiene que

$$
\begin{aligned}
m(m-1)\widehat{Var}(\hat{t}_{y,p}) &= \sum_{i=1}^{m}\left(\frac{y_{k_i}}{p_{k_i}} - \hat{t}_{y,p}\right)^2 \\
&= \sum_{i=1}^{m}\left(\frac{y_{k_i}^2}{p_{k_i}^2} - 2\hat{t}_{y,p}\frac{y_{k_i}}{p_{k_i}} + \hat{t}_{y,p}^2\right) \\
&= \sum_{i=1}^{m}\left(\frac{y_{k_i}^2}{p_{k_i}^2}\right) - 2\hat{t}_{y,p}\sum_{i=1}^{m}\frac{y_{k_i}}{p_{k_i}} + m\hat{t}_{y,p} \\
&= \sum_{i=1}^{m}\left(\frac{y_{k_i}^2}{p_{k_i}^2}\right) - 2m\hat{t}_{y,p}^2 + m\hat{t}_{y,p} \\
&= \sum_{i=1}^{m}\left(\frac{y_{k_i}}{p_{k_i}}\right)^2 - m\hat{t}_{y,p}^2
\end{aligned}
$$

■

Aunque el diseño muestral sea con reemplazo, es posible utilizar el estimador de Horvitz-Thompson, pues conserva su insesgamiento. La comparación entre la precisión del estimador de Horvitz-Thompson y el estimador de Hansen-Hurwitz, en un diseño con repetición depende de la configuración de los valores de la característica de interés en la población y_k $\forall k = 1, 2, ..., N$. Sin embargo, generalmente el estimador de Horvitz-Thompson es más eficiente más eficiente que el estimador de Hansen-Hurwitz, aunque éste último es más fácil de calcular. Cuando el diseño de muestreo es de tamaño fijo, el estimador de Horvitz-Thompson y Hansen-Hurwitz coinciden.

Ejemplo 2.2.3. Continuando con el ejercicio léxico-gráfico de la estimación del total poblacional t_y para todas las posibles muestras con reemplazo de tamaño 2 de la población U, tenemos la siguiente tabla que da cuenta del soporte de muestreo con ayuda de la función `SupportWR`

```
  X1       X2         p.k    p.l       p.s       y.k  y.l    Est
Yves     Yves       0.250  0.250    0.0625       32   32    128
Yves     Ken        0.250  0.250    0.1250       32   34    132
Yves     Erik       0.250  0.125    0.0625       32   46    248
Yves     Sharon     0.250  0.125    0.0625       32   89    420
Yves     Leslie     0.250  0.250    0.125        32   35    134
Ken      Ken        0.250  0.250    0.0625       34   34    136
Ken      Erik       0.250  0.125    0.0625       34   46    252
Ken      Sharon     0.250  0.125    0.0625       34   89    424
Ken      Leslie     0.250  0.250    0.125        34   35    138
Erik     Erik       0.125  0.125    0.015625     46   46    368
Erik     Sharon     0.125  0.125    0.03125      46   89    540
Erik     Leslie     0.125  0.250    0.0625       46   35    254
Sharon  Sharon      0.125  0.125    0.015625     89   89    712
Sharon   Leslie     0.125  0.250    0.0625       89   35    426
Leslie  Leslie      0.250  0.250    0.062500     35   35    140
```

El vector `Est` contiene las estimaciones de Hansen-Hurwitz para cada una de las posibles 15 muestras con reemplazo, su esperanza se calcula como

```
> sum(est*p.s)
[1] 236
```

Nótese que la esperanza del estimador equivale al total de la característica de interés, corroborando su insesgamiento. Por otro lado, para seleccionar una muestra con reemplazo, R incorpora la función `sample`, cuyos principales argumentos son

`x, size, replace, prob`.

`x` es el tamaño de la población, `size` es un número entero que determina el tamaño de la muestra. Para seleccionar una muestra con reemplazo, el argumento `replace` debe tomar el valor `TRUE`, así `replace = TRUE`. Cada elemento perteneciente a la población debe tener asociado un vector de probabilidades de selección cuya suma sea igual a la unidad. En R, el argumento `prob` contiene este vector de probabilidades; cuando se omite este argumento, la función `sample` asume que las probabilidades de selección son idénticas para cada individuo en la población. Así, por ejemplo, para seleccionar una muestra con reemplazo del marco de muestreo de U de tamaño $m = 3$, con las probabilidades de selección dadas por

```
> pk  <-  c(0.9, 0.025, 0.025, 0.025, 0.025)
> pk
[1] 0.900 0.025 0.025 0.025 0.025
```

Nótese que la suma de las probabilidades de selección es igual a uno y que los rótulos o nombres para cada individuo en la población están contenidos en el objeto `U`.

```
> U
[1] "Yves"   "Ken"    "Erik"   "Sharon" "Leslie"
```

Para seleccionar una muestra con reemplazo de tamaño $m = 3$ se debe escribir el siguiente código

```
> sam  <-  sample(N, 3, replace=TRUE, prob=pk)
> sam
[1] 3 1 1
```

Para la selección anterior, fue escogido dos veces el primer elemento y una vez el tercer elemento. La indexación de los rótulos (nombres) y valores de la característica de interés de los elementos escogidos en la muestra se hace utilizando

```
pkm <- c(0.9, 0.025, 0.025)
ym <- c(46, 32, 32)
```

Nótese que el tamaño de muestra es 3, pero el tamaño efectivo de muestra es $n(S) = 2$. Siendo `pkm` el vector de probabilidades de selección para los individuos pertenecientes a la muestra y `ym` el vector de valores de la característica de interés para los individuos pertenecientes a la muestra. La función `HH` del paquete `TeachingSampling` realiza la estimación del total poblacional para la característica de interés. Esta función consta de dos argumentos: `y`, el vector de valores de la característica de interés de los individuos en la muestra y `pk` sus correspondientes probabilidades de selección.

```
> est <- HH(ym,pkm)[1]
> est
         [,1]
[1,] 870.370
```

Para realizar la estimación de la varianza se crea un vector de diferencias `dif` entre $\frac{y_i}{p_i}$ y la estimación. Luego se procede a elevarlo al cuadrado, sumarlo y dividir por $m(m-1)$.

```
> dif <- rep(0,3)
> dif[1] <- (ym[1]/pkm[1])-est
> dif[2] <- (ym[2]/pkm[2])-est
> dif[3] <- (ym[3]/pkm[3])-est

> dif
[1] -819.2593  409.6296  409.6296
> Var <- (1/3)*(1/2)*sum(dif^2)
> Var
[1] 167796.4
```

Luego, el respectivo coeficiente de variación estimado es

$$cve(\hat{t}_p) = \frac{\sqrt{195018}}{637.037} \cong 0.7$$

Podemos pensar en el coeficiente de variación estimado como una medida de precisión. Así, las anteriores estimaciones se podrían decir inaceptables porque esta medida es muy alta.

El objetivo de este libro es que el lector esté en la capacidad de proponer estrategias de muestreo que permitan estimaciones precisas y confiables. Es decir, estimaciones cuyo coeficiente de variación sea aceptable[7] cuya longitud del intervalo de confianza sea corta con un nivel de confianza satisfactorio.

2.3 Muestras representativas

La teoría de muestreo se ha visto enriquecida en las últimas décadas por valiosos aportes a nivel mundial; aunque la base de la teoría de muestreo es la teoría

[7] En muchos casos un coeficiente de variación aceptable es menor al 3 por ciento.

de probabilidad, cuyo desarrollo axiomático cuenta varios centenares de años, su desarrollo práctico no sucedió sino hasta comienzos del siglo XX. Sin embargo, en la teoría clásica de inferencia estadística, basados en el pensamiento de Ronald Fisher y otros, asumen que la población es infinita. Un aspecto fundamental de la teoría de muestreo es que está basada en la realidad, en donde las poblaciones por más grandes que sean son de naturaleza finita.

Partiendo de este hecho es posible fundamentar la inferencia basada en una muestra aleatoria pero que proviene de una población finita y desde esta perspectiva los resultados de las inferencias diferirán de una manera significativa. De hecho, el llamado de atención es para que las personas que hacen inferencia con datos provenientes de un estudio por muestreo, se actualicen y no cometan grandes equivocaciones a la hora de presentar los resultados de la inferencia (Chambers & Skinner 2003). Por eso la teoría de muestreo cubre aspectos fundamentales de la estadística, porque desde un experimento controlado, hasta una encuesta por muestreo (Survey sampling), se debe pensar en el mecanismo de recolección de la información, y desde allí en la inferencia.

Un ejemplo común en las aulas de clase es describir la población en el tablero mediante una carita feliz, el profesor dice que una muestra representativa de la población es aquella muestra en donde se sigue viendo la misma carita feliz. Es decir, existe la creencia que una muestra representativa es un modelo reducido de la población y de aquí se desprende un argumento de validez sobre la muestra: una buena muestra es aquella que se parece a la población, de tal forma que las categorías aparecen con las mismas proporciones que en la población. Nada más falso que esta creencia. En algunos casos es fundamental sobre-representar algunas categorías o incluso seleccionar unidades con probabilidades desiguales.

Tillé (2006) cita el siguiente ejemplo: suponga que el objetivo es estimar la producción de hierro en un país y que nosotros sabemos que el hierro es producido, por dos compañías gigantes con miles de empleados y por cientos de pequeñas compañías con pocos empleados. ¿La mejor forma de seleccionar la muestra consiste en asignar la misma probabilidad a cada compañía? Claro que no. Primero averiguamos la producción de las grandes compañías. Después, seleccionamos una muestra de las compañías pequeñas.

La muestra no debe ser un modelo reducido de la población; debe ser una herramienta usada para obtener estimaciones. Es así como el concepto de muestra representativa pierde peso. Más aún, para Hájek (1981), una estrategia de muestreo es una dupla: diseño de muestreo (distribución de probabilidad sobre todas las posibles muestras) y estimador. La teoría de muestreo se ha ocupado de estudiar estrategias óptimas que permitan asegurar la calidad de las estimaciones. Entonces, el concepto de representatividad debería estar asociado con las estrategias de muestreo y no sólo con las muestras.

Siguiendo con Tillé (2006), una estrategia se dice representativa si permite estimar un total poblacional exactamente; es decir, sin sesgo y con varianza nula. Si se utiliza, por ejemplo, el estimador de Horvitz-Thompson junto con un diseño de muestreo apropiado, esta estrategia es representativa sólo sí, junto con la muestra seleccionada, el estimador reproduce algunos totales de la población; tales muestras se llaman muestras balanceadas. Existen también, estimadores que brindan a la estrategia el calificativo de representativa, algunos de ellos son conocidos como estimadores de calibración.

2.4 Ejercicios

2.1 Pruebe que bajo un diseño de muestreo $p(s)$, el error cuadrático medio de cualquier estimador $\hat{T}(s)$ de un parámetro T es igual a la varianza $Var(\hat{T})$ más el sesgo al cuadrado $B^2(\hat{T})$.

Sugerencia: $ECM\left(\hat{T}\right) = E_p\left(\hat{T}(s) - T\right)^2 = \sum_{s\in Q}\left(\hat{T}(s) - T\right)^2 p(s)$.

2.2 Demuestre que $\pi_{kl} = E_p\left(I_k(s)I_l(s)\right)$.

2.3 Suponga que tiene acceso a la población finita de tamaño $N = 5$ del ejemplo 2.2.1. y asuma el siguiente diseño de muestreo sin reemplazo

$$p(S = s) = \begin{cases} 0.2, & \text{para } s = \{Ken, Erik, Sharon\},\ s = \{Ken, Leslie\}, \\ 0.3, & \text{para } s = \{Yves, Erik, Leslie\},\ s = \{Yves, Sharon\}, \\ 0, & \text{En otro caso.} \end{cases}$$

- Calcule todas las probabilidades de inclusión de primer y de segundo orden.
- ¿Es el anterior un diseño de muestreo de tamaño de muestra fijo? Explique.
- Enumere todos los valores que toma la variable aleatoria $n(S)$ y verifique las relaciones $E_p(n(S)) = \sum_U \pi_k$ y $Var_p(n(S)) = \sum_U \pi_k - \left(\sum_U \pi_k\right)^2 + \sum\sum_{k\neq l}\pi_{kl}$.

2.4 Suponga que tiene acceso a la población finita de tamaño $N = 5$ del ejemplo 2.2.1. y asuma el siguiente diseño de muestreo sin reemplazo

$$p(S = s) = \begin{cases} 0.1, & \text{Si } n(S) = 3, \\ 0, & \text{En otro caso.} \end{cases}$$

- Defina todas las posibles muestras que pertenecen al soporte inducido por el anterior diseño de muestreo.
- Calcule todas las probabilidades de inclusión de primer y de segundo orden.
- Verifique que $\sum_U \pi_k = 3$ y que $\sum_U \pi_k - \left(\sum_U \pi_k\right)^2 + \sum\sum_{k\neq l}\pi_{kl} = 0$. Explique.
- Verifique que $\sum_U \pi_{k1} = 3 \times \pi_1$, $\sum_U \pi_{k2} = 3 \times \pi_2$, hasta $\sum_U \pi_{k5} = 3 \times \pi_5$.
- Calcule todas las posibles covarianzas Δ_{kl} y verifique que $\sum_U \Delta_{k1} = 0$, hasta $\sum_U \Delta_{k5} = 0$.

2.5 Demuestre o refute la siguiente afirmación: ≪Bajo cualquier diseño de muestreo, la suma poblacional de las probabilidades de inclusión de primer orden es siempre igual al tamaño de muestra≫.

2.6 Demuestre o refute la siguiente afirmación: ≪Bajo cualquier diseño de muestreo, el estimador de Horvitz-Thompson puede ser utilizado para obtener una estimación insesgada del total poblacional≫.

2.7 Suponga que tiene acceso a la población finita de tamaño $N = 5$ del ejemplo 2.2.1 y que y_k denota el valor de la característica de interés en el k-ésimo individuo. De esta manera, se tiene que:

$$y_{Yves} = 32, \quad y_{Ken} = 34, \quad y_{Erik} = 46, \quad y_{Sharon} = 89, \quad y_{Leslie} = 35$$

- Para el diseño de muestreo del ejercicio 2.3, en cada una de las posibles muestras calcule la estimación de Horvitz-Thompson, la estimación de la varianza, el *cve* y la estimación del intervalo de confianza al 95 %. Por último, muestre que el estimador es insesgado y calcule la varianza del estimador utilizando la expresión (2.2.4).
- Para el diseño de muestreo del ejercicio 2.4, en cada una de las posibles muestras calcule la estimación de Horvitz-Thompson, la estimación de la varianza, el *cve* y la estimación del intervalo de confianza al 95 %. Por último, muestre que el estimador es insesgado y calcule la varianza del estimador utilizando la expresión (2.2.4) y (2.2.5). ¿Son iguales estas varianzas? Explique.
- Para el diseño de muestreo del ejercicio 2.3, en cada una de las posibles muestras calcule la estimación de Horvitz-Thompson de la media (expresión 2.2.10), la estimación del tamaño poblacional (expresión 2.2.14), la estimación alternativa de la media (expresión 2.2.15) y la estimación alternativa del total (expresión 2.2.18).
- Para el diseño de muestreo del ejercicio 2.4, en cada una de las posibles muestras calcule la estimación de Horvitz-Thompson de la media (expresión 2.2.10), la estimación del tamaño poblacional (expresión 2.2.14), la estimación alternativa de la media (expresión 2.2.15) y la estimación alternativa del total (expresión 2.2.18).

2.8 Demuestre o refute la siguiente afirmación: «Bajo cualquier diseño de muestreo con reemplazo, el estimador de Hansen-Hurwitz puede ser utilizado para obtener una estimación insesgada del total poblacional».

2.9 Demuestre o refute la siguiente afirmación: «La probabilidad de selección de un individuo es siempre igual a su probabilidad de inclusión».

2.10 Demuestre o refute la siguiente afirmación: «Cualquier diseño de muestreo con reemplazo se puede ver com un caso particular de la distribución multinomial».

2.11 Demuestre o refute la siguiente afirmación: «Para una población de tamaño N, el número de posibles muestras con reemplazo de tamaño m es N^m».

2.12 Suponga que tiene acceso a la población finita de tamaño $N = 5$ de los anteriores ejercicios y asuma las siguientes probabilidades de selección

$$p_k = \begin{cases} 0.3, & \text{para } k = Yves, Leslie, \\ 0.2, & \text{para } k = Erik, \\ 0.1, & \text{para } k = Ken, Sharon. \end{cases}$$

- ¿Cuántas muestras con reemplazo de tamaño $m = 3$ se pueden seleccionar? Especifique explícitamente el diseño de muestreo para estas muestras y compruebe que $\sum_{s \in Q} p(s) = 1$.

- Para este diseño de muestreo, y teniendo en cuenta los valores de la característica de interés del ejercicio 2.7, en cada una de las posibles muestras calcule la estimación de Hansen-Hurwitz, la estimación de la varianza, el *cve* y la estimación del intervalo de confianza al 95 %. Por último, muestre que el estimador es insesgado y calcule la varianza del estimador utilizando la expresión (2.2.35).
- ¿Es posible utilizar otro tipo de estimadores para obtener estimaciones insesgadas del total poblacional?

2.13 Demuestre rigurosamente que el estimador de la varianza del estimador de Hansen-Hurwitz corresponde a la expresión (2.2.36).

Capítulo 3

Muestras con probabilidades simples

> Las muestras no están dadas, las muestras deben ser seleccionadas, asignadas o capturadas. El tamaño de la muestra no siempre es fijo. En estudios por muestreo, el tamaño de muestra es casi siempre una variable aleatoria. Los datos no siempre son independientes o idénticamente distribuidos y usualmente no son seleccionados de una sola población, sino de subpoblaciones compuestas o complementarias. Más aún, no se produce una sola estimación, se produce un conjunto de estimaciones, así que la historia que siempre nos han contado está equivocada.
>
> Leslie Kish (1996)

Cuando el marco de muestreo disponible para la selección de la muestra es una lista conteniendo la identificación y la ubicación de los elementos en la población, se utilizan diseños de muestreo que permitan la inclusión de éstos en la muestra de forma directa. Es decir, en la selección de la muestra, los elementos poblacionales son las mismas unidades de muestreo. Una vez que el procedimiento de muestreo ha seleccionado la muestra de elemento, el siguiente paso a realizar es la medición de la característica de interés y_k en cada elemento de la muestra seleccionada ($k \in s$).

En este capítulo se describen los diseños de muestreo para elementos más importantes, algunos de los cuales son ampliamente utilizados en la práctica, otros tienen la característica de ser de tamaño de muestra variable o aleatorio. Cuando el marco de muestreo contiene información auxiliar de tipo continuo para cada elemento de la población, se utilizará esta información en la selección de la muestra, induciendo los diseños proporcionales al tamaño. Cuando el marco de muestreo contiene información auxiliar discreta, se utilizarán diseños de muestra estratificados que permiten, a menudo, mayor precisión cuando la característica de interés presenta comportamientos diferentes en cada estrato o grupo poblacional.

Para cada diseño de muestreo se realiza una descripción teórica, se utilizará la población U para realizar algunos ejercicios léxico-gráficos que describan el comportamiento de la estrategia de muestreo. Por otro lado, se utilizará la población Lucy

y, con ayuda del paquete `TeachingSampling`, se seleccionará una única muestra para la posterior estimación de los parámetros de interés. También habrá ejemplos prácticos de la vida real que permiten una mayor comprensión de las características del diseño y un mayor conocimiento a la hora de decidir qué diseño de muestreo debe see implementado en determinados casos.

Las estrategias de muestreo implementadas en este capítulo corresponden a la utilización del estimador de Horvitz-Thompson junto con diseños de muestreo sin reemplazo y/o al uso del estimador de Hansen-Hurwitz en diseños de muestra con reemplazo.

3.1 Diseño de muestreo Bernoulli

En el diseño de muestreo Bernoulli se fija a priori (por experiencia o alguna otra razón) la probabilidad de inclusión de todos los individuos, la cual permanece constante para todo el universo. Es decir, $\pi_k = \pi$ para todo $k \in U$. Un típico ejemplo de la implementación de este diseño en la práctica es la revisión de equipajes de pasajeros por los funcionarios de la aduana en un aeropuerto; se fija la probabilidad de inclusión para cada pasajero y mediante cierto mecanismo de selección (muy simple) se selecciona la muestra, conforme las personas van ingresando al sitio. Nótese que el tamaño de muestra $n(S)$ es aleatorio porque una muestra realizada mediante este mecanismo de selección puede incluir a todos los pasajeros o a ningún pasajero de la población.

Definición 3.1.1. *Siendo $n(s)$ el tamaño de muestra, el diseño de muestreo Bernoulli selecciona la muestra s con probabilidad*

$$p(s) = \begin{cases} \pi^{n(s)}(1-\pi)^{N-n(s)} & \text{si } s \text{ tiene tamaño igual a } n(s) \\ 0 & \text{en otro caso} \end{cases} \tag{3.1.1}$$

3.1.1 Algoritmo de selección

La selección de una muestra con diseño Bernoulli conlleva los siguientes pasos:

1. Fijar el valor de π tal que $0 < \pi < 1$.
2. Obtener ε_k para $k \in U$ como N realizaciones independientes de una variable aleatoria con distribución uniforme sobre el intervalo $[0, 1]$.
3. El elemento k-ésimo pertenece a la muestra con probabilidad π. Es decir, si $\varepsilon_k < \pi$ el individuo k-ésimo es seleccionado.

Dado que $\varepsilon_k \sim Unif[0, 1]$, se tiene que $Pr(\varepsilon_k < \pi) = \pi$ para $k \in U$. Por tanto, la inclusión de los individuos k-ésimo y l-ésimo, para $k \neq l$, es independiente. Esto implica que la distribución de $I_k(S)$ es Bernoulli $Ber(\pi)$ y se tiene el siguiente resultado.

Resultado 3.1.1. *Definiendo a Q_r como el soporte que contiene a todas las posibles muestras de tamaño r, existen $\binom{N}{r}$ muestras pertenecientes a Q_r. En otras palabras*

$$\#(Q_r) = \binom{N}{r} \qquad r = 0, \ldots, N$$

Sin embargo, al definir Q como el soporte general de todas las posibles muestras de tamaños entre $r = 0$ y $r = N$, se tiene que

$$\#(Q) = \sum_{r=1}^{N} \binom{N}{r} = 2^N$$

Resultado 3.1.2. *Bajo muestreo Bernoulli, la distribución del tamaño de muestra $n(S)$ es binomial $Bin(N, \pi)$ y*

$$Pr(n(S) = r) = \sum_{s \in Q_r} p(s) = \binom{N}{r} \pi^r (1 - \pi)^{N-r}, \tag{3.1.2}$$

con $r = 1, \ldots, N$ y Q_r el soporte que contiene a todas las posibles muestras de tamaño r, donde $Q_r \subset Q$.

Prueba. La distribución de $I_k(S)$ es Bernoulli $Ber(\pi)$, las inclusiones de los individuos en la muestra son eventos independientes, entonces $n(S) = \sum_U I_k$ sigue una distribución binomial. Ahora, dado el diseño de muestreo (3.1.1), para cualquier $s \in Q_r$, se cumple que $p(s) = \pi^r(1 - \pi)^{N-r}$. Como existen $\binom{N}{r}$ maneras de seleccionar una muestra de r elementos de una población de tamaño N, se tiene que $\#(Q_r) = \binom{N}{r}$. Luego, al sumar $p(s)$ sobre todas las muestras del soporte Q_r se obtiene el resultado. ■

Como $n(S)$ es aleatorio, existen 2^N posibles muestras en el soporte Q. Nótese que $n(S)$ tiene una distribución Binomial y, por tanto, su esperanza y varianza están dadas por:

$$E(n(S)) = N\pi \qquad Var(n(S)) = N(\pi)(1 - \pi), \tag{3.1.3}$$

Aunque el investigador haya fijado las probabilidades de inclusión, se puede verificar que realmente el diseño de muestreo Bernoulli cumple las condiciones establecidas en el capítulo anterior y también que las probabilidades de inclusión, inducidas por el diseño de muestreo, son idénticas para cada elemento en la población $\pi_k = \pi$.

Resultado 3.1.3. *Bajo el diseño de muestreo Bernoulli, se verifica que*

$$\sum_{s \in Q} p(s) = 1 \tag{3.1.4}$$

Prueba. Para una población de tamaño N, el tamaño de muestra puede ser r con $r = 0, 1, \ldots, N$. Es suficiente probar que $\sum_{r=0}^{N} Pr(n(S) = r) = 1$, utilizando el teorema binomial se tiene de inmediato porque $n(S) \sim Bin(N, \pi)$. Más aún, se

tiene que

$$\begin{aligned}\sum_{s\in Q} p(s) &= \sum_{s\in Q_0} p(s) + \sum_{s\in Q_1} p(s) + \cdots + \sum_{s\in Q_N} p(s)\\ &= \binom{N}{0}\pi^0(1-\pi)^{N-0} + \cdots + \binom{N}{N}\pi^N(1-\pi)^{N-N}\\ &= \sum_{r=0}^{N}\binom{N}{r}\pi^r(1-\pi)^{N-r} = (\pi+1-\pi)^N = 1\end{aligned}$$

■

Resultado 3.1.4. *Para el diseño de muestreo Bernoulli, las probabilidades de inclusión de primer y segundo orden están dadas por:*

$$\pi_k = \pi \tag{3.1.5}$$

$$\pi_{kl} = \begin{cases}\pi & para k = l\\ \pi^2 & Para\ k \neq l\end{cases} \tag{3.1.6}$$

Prueba. Teniendo en cuenta que existen $\binom{N-1}{r-1}$ muestras de tamaño r que contienen al elemento k-ésimo, tenemos

$$\begin{aligned}\pi_k &= \sum_{\substack{s\ni k\\ s\subset Q}} p(s)\\ &= \sum_{\substack{s\ni k\\ s\subset Q_0}} p(s) + \sum_{\substack{s\ni k\\ s\subset Q_1}} p(s) + \cdots + \sum_{\substack{s\ni k\\ s\subset Q_N}} p(s)\\ &= 0 + \binom{N-1}{0}\pi(1-\pi)^{N-1} + \ldots + \binom{N-1}{N-1}\pi(1-\pi)^{N-1}\\ &= \sum_{r=0}^{N-1}\binom{N-1}{r}\pi^{r+1}(1-\pi)^{N-1-r}\\ &= \pi\sum_{r=0}^{N-1}\binom{N-1}{r}\pi^r(1-\pi)^{N-1-r} = \pi(\pi+(1-\pi))^{N-1} = \pi\end{aligned}$$

Donde se utiliza el resultado del teorema binomial (Mood, Graybill & Boes 1974) que afirma que

$$\sum_{r=0}^{m}\binom{m}{r}a^r b^{m-r} = (a+b)^m. \tag{3.1.7}$$

Ahora como las inclusiones de los elementos de la población en la muestra son eventos independientes, entonces

$$Pr(k\in S \text{ y } l\in S) = Pr(I_k = 1)Pr(I_l = 1) = \pi^2 \tag{3.1.8}$$

■

3.1.2 El estimador de Horvitz-Thompson

Resultado 3.1.5. *Para el diseño de muestreo Bernoulli, el estimador de Horvitz-Thompson, su varianza y su varianza estimada están dados por:*

$$\hat{t}_{y,\pi} = \frac{1}{\pi} \sum_S y_k \tag{3.1.9}$$

$$Var_{BER}(\hat{t}_{y,\pi}) = \left(\frac{1}{\pi} - 1\right) \sum_U y_k^2 \tag{3.1.10}$$

$$\widehat{Var}_{BER}(\hat{t}_{y,\pi}) = \frac{1}{\pi}\left(\frac{1}{\pi} - 1\right) \sum_S y_k^2, \tag{3.1.11}$$

respectivamente

Prueba. El resultado es inmediato porque

$$\Delta_{kl} = \begin{cases} \pi_{kl} - \pi_k \pi_l = \pi^2 - \pi^2) = 0 & \text{para } k \neq l \\ \pi_{kk} - \pi_k \pi_k = \pi(1-\pi) & \text{para } k = l \end{cases} \tag{3.1.12}$$

luego la doble suma en la varianza del estimador de Horvitz-Thompson pasa a ser una sola suma; lo anterior sucede análogamente con la expresión de la estimación de la varianza. ■

Nótese que en caso de que la muestra realizada o seleccionada esté compuesta por todas las unidades de la población, es decir se deba realizar un censo[1], la probabilidad de inclusión para cada elemento de la población estaría dada por $\pi_k = \pi$. En este caso, el estimador de Horvitz-Thompson estaría dado por la siguiente expresión

$$\hat{t}_{y,\pi} = \frac{1}{\pi} \sum_U y_k = \frac{t_y}{\pi} \neq t_y \tag{3.1.13}$$

En este caso, el estimador de Horvitz-Thompson es deficiente para la estimación del total poblacional t_y y se sugiere la utilización del estimador alternativo para el total poblacional que, para el caso particular del diseño de muestreo Bernoulli, estaría dado por

$$\hat{t}_{y,alt} = N\widetilde{y}_S = N\frac{\sum_S y_k}{n(S)} = N\bar{y}_S. \tag{3.1.14}$$

Fácilmente se verifica que si $s = U$, entonces $\hat{t}_{y,alt} = t_y$.

Ejemplo 3.1.1. Para nuestra población de ejemplo U, existen $2^5 = 32$ posibles muestras. Si la probabilidad de inclusión es fija para cada elemento e igual a 0,3, realice el cálculo léxico-gráfico del estimador de Horvitz-Thompson y compruebe el insesgamiento y la varianza.

[1]En el diseño de muestreo Bernoulli, la probabilidad de seleccionar todas las unidades de la población en la muestra es equivalente a π^N.

3.1.3 Marco y Lucy

Suponga que se debe seleccionar una muestra con un diseño de muestreo Bernoulli. Se quiere que el tamaño esperado de muestra sea de $N\pi = 400$ empresas del sector industrial. Como el tamaño de la población es $N = 2396$, entonces el valor que se fija para π es de 0.1669. Para seleccionar la muestra se utiliza la función `S.BE(N,prob)` del paquete `TeachingSampling` cuyos parámetros son `N`, el tamaño poblacional y `prob` el valor de la probabilidad de inclusión para cada elemento de la población. Esta función utiliza el algoritmo secuencial descrito en la anterior sección.

Primero se carga en R el archivo `Marco` que contiene el marco de muestreo para la selección de la muestra. Se fijan los parámetros de la función, `N` y `prob`. Esta función devuelve un vector conteniendo el índice de los elementos seleccionados en la muestra. En este caso particular, el primer elemento seleccionado es el número 2 y el último el número 2394.

```
> data(Lucy)
> attach(Lucy)
> N <- dim(Lucy)[1]
> pik <- 0.1669
> sam <- S.BE(N,pik)
> muestra <- Lucy[sam,]
> attach(muestra)
> muestra
          ID Ubication  Level Zone
2     AB002       c1k2  Small    A
8     AB008       c1k8  Small    A
10    AB010      c1k10  Small    A
14    AB014      c1k14  Small    A
...
2392  AB984     c26k16    Big    E
2394  AB986     c26k18    Big    E
> n <- dim(muestra)[1]
> n
[1] 396
```

Aplicando los índices obtenidos por la función `S.BE` al marco de muestreo obtenemos la identificación y ubicación de las empresas seleccionadas en la muestra. Nótese que el tamaño de muestra efectivo es de 396 empresas. Una vez que la etapa de recolección de datos se haya realizado, obtendremos un archivo de datos de `Lucy` conteniendo los valores de las características de interés para las empresas seleccionadas que será adjuntado a R mediante la función `attach`.

La etapa de estimación de resultados se hace utilizando la función `E.BE(y,prob)` del paquete `TeachingSampling` cuyos argumentos son `y`, un vector o matriz conteniendo los valores de las características de interés en la muestra y `prob`, la probabilidad de inclusión. En este caso la longitud de cada vector es de $n = 396$. Esta función arroja la estimación del total poblacional de `y` usando el estimador de Horvitz-Thompson, la estimación de la varianza y el coeficiente de variación del mismo. Por ejemplo, la variable `Income` contiene los valores del ingreso declarado

en el último año por 396 empresas del sector industrial pertenecientes a la muestra. La estimación para esta característica se hace mediante el siguiente código:

```
> estima <- data.frame(Income, Employees, Taxes)
> E.BE(estima,pik)
                   Income    Employees        Taxes
Estimation 1.003565e+06 1.450449e+05 2.898143e+04
Variance   3.141990e+09 5.862754e+07 5.062942e+06
CVE        5.585434e+00 5.278959e+00 7.763932e+00
```

La tabla 3.1. muestra los resultados obtenidos para este caso particular, donde la desviación relativa de una estimación, medida en porcentaje está definida como

$$DR = 100 \times \frac{\hat{t}_{y,\pi} - t_y}{t_y} \tag{3.1.15}$$

Tabla 3.1: *Muestreo Bernoulli: estimación de los totales de las características de interés.*

Variable	Total poblacional	Total estimado	cve %	Desv. %
Ingreso	1035217	1003565	5.58	-3.06
Empleados	151950	145044	5.27	-4.55
Impuestos	28654	28981	7.76	1.14

Nótese que, aunque la distribución asintótica del estimador de Horvitz-Thompson es normal, es necesario verificar el comportamiento del estimador con el tamaño de muestra esperado. Se realizaron varios experimentos de Monte Carlo con el propósito de tener un examen más cercano del estimador de Horvitz-Thompson del total de la característica `Income` en la población `Lucy`. El resultado de la simulación se muestra en los histogramas de la figura 3.1. Se espera que el promedio de las estimaciones en cada experimento coincida con el total poblacional y la varianza de éstas debe acercarse a la varianza basada en el diseño de muestreo Bernoulli.

La media de las estimaciones de t_y es 1035176 que ajusta bien con el parámetro correspondiente $t_y = 1035217$. La distribución parece ser simétrica con forma de campana (los valores de la distribución teórica se muestran en la curva sólida y roja) y no se notan grandes discrepancias entre lo observado y lo teórico. En algunos casos, en donde el tamaño de muestra no es lo suficientemente grande, se debe verificar el comportamiento normal del estimador.

3.2 Muestreo aleatorio simple sin reemplazo

El muestreo aleatorio simple puede ser visto como la forma más básica de selección de muestras. Supone la existencia de homogeneidad en los valores poblacionales de la característica de interés. Partiendo de esta asunción, este diseño provee probabilidades de selección idénticas para cada una de las posibles muestras pertenecientes

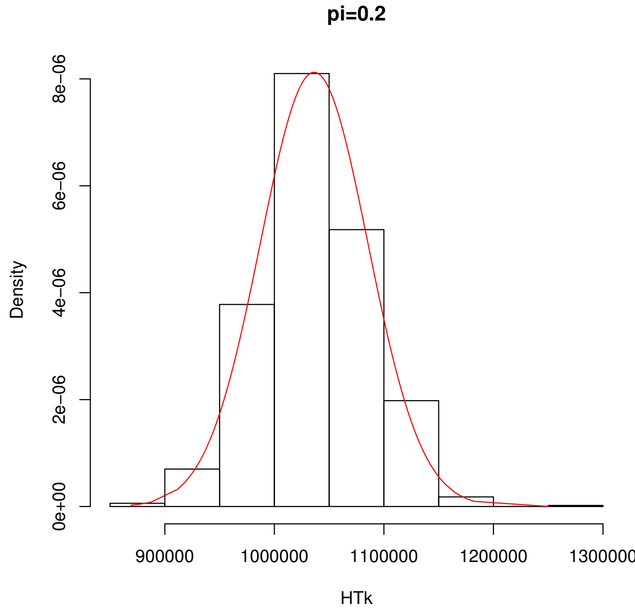


Figura 3.1: *Distribución muestral del estimador de Horvitz-Thompson bajo diseño Bernoulli.*

al soporte Q. Lohr (2000) cita un ejemplo al respecto del uso del diseño de muestreo aleatorio simple diciendo que, cuando la población es homogénea, el investigador no necesita examinar todos los elementos de la población así como el encargado del análisis médico no necesita obtener toda la sangre para medir la cantidad de glóbulos rojos.

Una **muestra aleatoria simple sin reemplazo** de tamaño n se elige de modo que cada posible muestra realizada de tamaño n tenga la misma probabilidad de ser seleccionada. A diferencia del diseño de muestreo Bernoulli, el diseño de muestreo aleatorio simple sin reemplazo tiene la característica de ser de tamaño fijo. Una **muestra aleatoria simple con reemplazo**, de tamaño m de una población de N elementos es la extracción de m muestras independientes de tamaño 1, en donde cada elemento se extrae de la población con la misma probabilidad.

Lehtonen & Pahkinen (2003) afirman que este diseño de muestreo no es muy común en la práctica y básicamente desempeña dos funciones. Primero, plantean una línea de comparación de la eficiencia relativa con otros diseños de muestreo. Segundo, dentro de los diseños de muestreo más sofisticados como diseños de muestreo estratificado o diseños de muestreo por conglomerados, el muestreo aleatorio simple puede ser utilizado como un método final de selección de unidades primarias.

Definición 3.2.1. *Un diseño de muestreo se dice aleatorio simple sin reemplazo si todas las posibles muestras de tamaño n tienen la misma probabilidad de ser seleccionadas. Así,*

$$p(s) = \begin{cases} \frac{1}{\binom{N}{n}} & \text{si } \#s = n \\ 0 & \text{en otro caso} \end{cases} \tag{3.2.1}$$

Resultado 3.2.1. *Definiendo a Q como el soporte que contiene a todas las posibles muestras de tamaño n, existen $\binom{N}{n}$ muestras pertenecientes a Q. En otras palabras,*

$$\#(Q) = \binom{N}{n}$$

Nótese que $\sum_{s\in Q} p(s) = 1$ porque $\#Q = \binom{N}{n}$.

3.2.1 Algoritmos de selección

Durante muchos años, la teoría de muestreo se centró en la parte de la extracción de muestras aleatorias, más que en la construcción de los estimadores. Con la gran ventaja de los nuevos procesadores, lo anterior pasa a un segundo plano. A continuación se presentan dos métodos de selección de una muestra aleatoria simple de tamaño n de una población de tamaño N. Existen bastantes métodos de selección de una muestra aleatoria sin reemplazo, en esta sección se abordan dos algoritmos de selección. El primero da una asunción más simple, y puede ser comparado con el conocido método de la extracción de una balota; sin embargo, Tillé (2006) afirma que este método es ineficiente computacionalmente. El segundo método basado en un algoritmo secuencial, permite la selección de la muestra con una sola revisión del marco de muestreo.

Método coordinado negativo

Sunter (1977) ha probado que el siguiente método de ordenamiento aleatorio arroja como resultado una muestra aleatoria simple. Para extraer la muestra de tamaño n de un universo de N objetos,

1. Generar N realizaciones de una variable aleatoria ξ_k $(k \in U)$ con distribución uniforme (0,1).
2. Asignar ξ_k al elemento k-ésimo de la población.
3. Ordenar la lista de elementos descendente (o ascendentemente) con respecto a este número aleatorio ξ_k.
4. A continuación, seleccionar los n primeros (o los n últimos) elementos. Esta selección corresponde a la muestra realizada.

Es necesario tener la seguridad de que exista un número grande de décimas en cada ξ_k para evitar problemas de empates (números aleatorios repetidos).

Método de selección y rechazo

Fan, Muller & Rezucha (1962) implementaron el siguiente algoritmo de muestreo secuencial (porque se recorre el marco de muestreo, elemento por elemento, y se decide la pertenencia o el rechazo del objeto en la muestra). Es interesante que, más tarde Bebbington (1975) trece años más tarde publica (en un artículo de una página) el mismo método, aunque sin escribir ninguna fórmula.

En general se supone que el marco de muestreo tiene N individuos, y se quiere seleccionar una muestra aleatoria de n individuos. Así, para el individuo k ($k = 1, 2, ..., N$), se tiene que

1. Realizar $\xi_k \sim U(0, 1)$
2. Calcular
$$c_k = \frac{n - n_k}{N - k + 1}$$
donde n_k es la cantidad de objetos seleccionados en los $k - 1$ ensayos anteriores.
3. Si $\xi_k < c_k$, entonces el elemento k pertenece a la muestra.
4. Detener el proceso cuando $n = n_k$.

Dado que este algoritmo se detiene cuando $n = n_k$, resulta muy eficiente porque asegura una muestra aleatoria simple y en algunas ocasiones no se requiere recorrer todo el marco de muestreo.

Ejemplo 3.2.1. Para seleccionar muestras aleatorias simples, R incorpora la función `sample`. Ésta, por defecto selecciona muestras sin reemplazo. Así, por ejemplo, para seleccionar una muestra aleatoria de tamaño $n = 2$, de la población de ejemplo `U` de tamaño $N = 5$, sin reemplazo se tiene

```
> sam <- sample(N, 2, replace=FALSE)
> U[sam]
[1] "Ken"    "Leslie"
```

El algoritmo de selección y rechazo está implementado en la función `S.SI` del paquete `TeachingSampling` cuyos argumentos son el tamaño de la población `N`, el tamaño de muestra deseado `n` y un vector de números aleatorios `e` que, por defecto, se asigna mediante la generación de `N` realizaciones de una variable aleatoria con distribución uniforme en el intervalo $]0, 1[$.

Para seleccionar una muestra aleatoria sin reemplazo de tamaño $n = 2$ por el método de selección y rechazo, de la población de ejemplo `U` de tamaño $N = 5$, sólo basta digitar el siguiente código.

```
> sam <- S.SI(N, 2)
> U[sam]
[1] "Erik"    "Leslie"
```

Nótese que el resultado de la función `S.SI` es un vector de índices, que aplicados al identificador resulta en una muestra seleccionada que está conformada por los elementos **Erik** y **Leslie**.

La siguiente salida muestra cada uno de los `N=5` pasos del algoritmo. Los números aleatorios que se utilizaron están en la columna llamada `ek` y los índices de la muestra seleccionada están en la columna `sam`.

k	Nombre	ek	ck	nk	sam
1	Yves	0.4938	0.4000000	0	0
2	Ken	0.7044	0.5000000	0	0
3	Erik	0.4585	0.6666667	1	3
4	Sharon	0.6747	0.5000000	1	0
5	Leslie	0.8565	1.0000000	2	5

Resultado 3.2.2. *El diseño de muestreo Bernoulli coincide con el diseño de muestreo aleatorio simple sin reemplazo cuando el tamaño de muestra se considera fijo e igual a n.*

Prueba. Utilizando las propiedades de la probabilidad condicional se tiene que

$$\begin{aligned} Pr(S=s|n(S)=n) &= \frac{Pr(S=s \text{ y } n(S)=n)}{Pr(n(S)=n)} \\ &= \frac{\pi^n(1-\pi)^{N-n}}{\binom{N}{n}\pi^n(1-\pi)^{N-n}} = \frac{1}{\binom{N}{n}} \end{aligned}$$

el cual coincide con la expresión (3.2.1). ■

Una consecuencia inmediata del anterior resultado es que otro método de selección de muestras para un diseño de muestreo Bernoulli es escoger aleatoriamente el tamaño de muestra de acuerdo a una distribución binomial $Bin(N, \pi)$ y luego seleccionar una muestra mediante uno de los anteriores algoritmos de selección de muestras aleatorias simples sin reemplazo (Tillé 2006).

3.2.2 El estimador de Horvitz-Thompson

Resultado 3.2.3. *Para un diseño de muestreo aleatorio simple, las probabilidades de inclusión de primer y segundo orden están dadas por:*

$$\pi_k = \frac{n}{N} \tag{3.2.2}$$

$$\pi_{kl} = \frac{n(n-1)}{N(N-1)} \tag{3.2.3}$$

respectivamente. La covarianza de las variables indicadoras está dada por

$$\Delta_{kl} = \begin{cases} \pi_{kl} - \pi_k\pi_l = -\frac{n}{N^2}\frac{(N-n)}{(N-1)} & \text{para } k \neq l \\ \pi_k(1-\pi_k) = \frac{n(N-n)}{N^2} & \text{para } k = l \end{cases} \tag{3.2.4}$$

Prueba. Recurriendo a la definición de probabilidad de inclusión de primer orden, se tiene que

$$\begin{aligned} \pi_k &= Pr(I_k(S)=1) \\ &= \frac{\binom{1}{1}\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N} \end{aligned}$$

por otro lado,

$$
\begin{aligned}
\pi_k l &= Pr(k \in S \text{ y } l \in s)\\
&= Pr(I_k(S) = 1 \text{ y } I_l(S) = 1)\\
&= Pr(I_k(S) = 1 | I_l(S) = 1) Pr(I_l(s) = 1)\\
&= \frac{n-1}{N-1}\frac{n}{N} = \frac{n(n-1)}{N(N-1)}
\end{aligned}
$$

■

Resultado 3.2.4. *Para un diseño de muestreo aleatorio simple, el estimador de Horvitz-Thompson del total poblacional t_y, su varianza y su varianza estimada están dados por:*

$$\hat{t}_{y,\pi} = \frac{N}{n}\sum_S y_k \tag{3.2.5}$$

$$Var_{MAS}(\hat{t}_{y,\pi}) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right) S^2_{yU} \tag{3.2.6}$$

$$\widehat{Var}_{MAS}(\hat{t}_{y,\pi}) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right) S^2_{yS} \tag{3.2.7}$$

respectivamente, con

$$S^2_{yU} = \frac{1}{N-1}\sum_{k\in U}(y_k - \bar{y}_U)^2, \tag{3.2.8}$$

la ***varianza poblacional*** *de la característica de interés en el universo U y con*

$$S^2_{yS} = \frac{1}{n-1}\sum_{k\in S}(y_k - \bar{y}_S)^2 \tag{3.2.9}$$

la ***varianza muestral*** *de los valores de la característica de interés en la muestra aleatoria S. Además, $\bar{y}_S = \frac{\sum_S y_k}{n}$. Por otro lado, nótese que $\hat{t}_{y,\pi}$ es insesgado para el total poblacional t_y de la característica de interés y, y que $\widehat{Var}_{MAS}(\hat{t}_{y,\pi})$ es insesgado para $Var_{MAS}(\hat{t}_{y,\pi})$.*

Prueba. Por el resultado anterior, tenemos

$$\hat{t}_{y,\pi} = \sum_S \frac{y_k}{\pi_k} = \frac{N}{n}\sum_S y_k. \tag{3.2.10}$$

La demostración de las varianzas es inmediata al reemplazar las cantidades apropiadas en la expresión genérica del capítulo anterior y teniendo en cuenta que

$$\sum\sum_{k\neq l} y_k y_l = \sum_k\sum_l y_k y_l - \sum\sum_{k=l} y_k y_l = \left(\sum_U y_k\right)^2 - \sum_U y_k^2$$

De tal forma que,

$$
\begin{aligned}
Var(\hat{t}_{y,\pi}) &= \frac{N^2}{n^2} Var\left(\sum_U I_k(s) y_k\right) \\
&= \frac{N^2}{n^2}\left(\sum_U Var(I_k(s)) y_k^2 + \sum\sum_{k\neq l} Cov\left(I_k(S), I_l(s)\right) y_k y_l\right) \\
&= \frac{N^2}{n^2}\left(\frac{n(N-n)}{N^2}\sum_U y_k^2 - \frac{n}{N^2}\frac{(N-n)}{(N-1)}\sum\sum_{k\neq l} y_k y_l\right) \\
&= \frac{(N-n)}{n}\left(\sum_U y_k^2 - \frac{1}{N-1}\sum\sum_{k\neq l} y_k y_l\right) \\
&= \frac{(N-n)}{n}\frac{1}{N-1}\left((N-1)\sum_U y_k^2 - \left[\left(\sum_U y_k\right)^2 - \sum_U y_k^2\right]\right) \\
&= \frac{N(N-n)}{n}\frac{1}{N-1}\left(\sum_U y_k^2 - \frac{\left(\sum_U y_k\right)^2}{N}\right) \\
&= \frac{N^2}{n}\left(1-\frac{n}{N}\right) S_{yU}^2
\end{aligned}
$$

Para demostrar el insesgamiento de la varianza estimada es suficiente demostrar que S_{ys}^2 es insesgado para S_{yU}^2.

$$
\begin{aligned}
E(S_{yS}^2) &= E\left(\frac{1}{n-1}\left[\sum_S y_k^2 - n\bar{y}_S^2\right]\right) \\
&= \frac{1}{n-1}\left(E\left[\sum_S y_k^2\right] - nE\left[\frac{\hat{t}_{y,\pi}}{N}\right]^2\right) \\
&= \frac{1}{n-1}\left(\frac{n}{N}\left[\sum_U y_k^2\right] - \frac{n}{N^2}E\left[\hat{t}_{y,\pi}\right]^2\right) \\
&= \frac{1}{n-1}\left(\frac{n}{N}\left[\sum_U y_k^2\right] - \frac{n}{N^2}\left[\frac{N^2}{n}\left(1-\frac{n}{N}\right)S_{yU}^2 - t_y^2\right]\right) \\
&= \frac{n}{n-1}\left(\frac{1}{N}\left[\sum_U y_k^2\right] - \frac{1}{n}\left(1-\frac{n}{N}\right)S_{yU}^2 - \frac{t_y^2}{N^2}\right) \\
&= \frac{n}{n-1}\left(\frac{N-1}{N}S_{yU}^2 - \frac{N-n}{nN}S_{yU}^2\right) \\
&= S_{yU}^2
\end{aligned}
$$

En donde se utilizó el hecho de que $\bar{y}_S = \frac{\hat{t}_{y,\pi}}{N}$ y además

$$E(\hat{t}_{y,\pi})^2 = Var(\hat{t}_{y,\pi}) - t_y^2.$$

■

Ejemplo 3.2.2. Para nuestra población de ejemplo U, existen $\binom{5}{2} = 10$ posibles muestras de tamaño $n = 2$. Realice el cálculo léxico-gráfico del estimador de Horvitz-Thompson y compruebe el insesgamiento y la varianza.

3.2.3 Estimación de la media poblacional

Resultado 3.2.5. *Para un diseño de muestreo aleatorio simple, el estimador de Horvitz-Thompson para la media poblacional $\bar{y}_U$, su varianza y su varianza estimada están dados por:*

$$\hat{\bar{y}}_{\pi} = \frac{\hat{t}_{y,\pi}}{N} = \frac{\sum_S y_k}{n} = \bar{y}_S \tag{3.2.11}$$

$$Var_{MAS}(\hat{\bar{y}}_{\pi}) = \frac{1}{N^2} Var(\hat{t}_{y,\pi}) = \left(1 - \frac{n}{N}\right) \frac{S^2_{yU}}{n} \tag{3.2.12}$$

$$\widehat{Var}_{MAS}(\hat{\bar{y}}_{\pi}) = \frac{1}{N^2} Var(\hat{t}_{y,\pi}) = \left(1 - \frac{n}{N}\right) \frac{S^2_{ys}}{n} \tag{3.2.13}$$

respectivamente, con S^2_{yU} y S^2_{ys} el estimador de la varianza de los valores de la característica de interés y en el universo y en la muestra. Nótese que $\hat{t}_{y,\pi}$ es insesgado para el total poblacional t_y de la característica de interés y, y que $\widehat{Var}_{MAS}(\hat{t}_{y,\pi})$ es insesgado para $Var_{MAS}(\hat{t}_{y,\pi})$.

Nótese que la construcción, cálculo y estimación de la varianza son muy intuitivas. Haciendo un símil con la inferencia clásica, suponga que tenemos una muestra aleatoria $X_1, \ldots, X_n$ i.i.d., tal que $X_i \sim (\mu, \sigma^2)$. Se sabe que un estimador insesgado para la media μ es $\bar{X}$, además se sabe que la variación de este estimador es $\dfrac{\sigma^2}{n}$.

Al operador $\left(1 - \dfrac{n}{N}\right)$ se le conoce con el nombre de **factor de corrección para poblaciones finitas**. Sólo existe una sola muestra que contiene a todos los elementos de la población, por tanto, si esa muestra es seleccionada, esperamos que no haya variación en el estimador pues reproducirá con exactitud al parámetro, por tanto la varianza del mismo se debe anular. Entre más grande sea el tamaño de muestra n, al utilizar un diseño de muestreo aleatorio simple, la variabilidad de las estimaciones se debe hacer más pequeña dado que la muestra tenderá a parecerse más a la población finita. Lohr (2000) afirma que el tamaño de muestra es el que determina la precisión de las estimaciones (no así, el porcentaje de la población muestreada):

> Si su sopa está bien revuelta, sólo necesita dos o tres cucharadas para probar el sazón, así tenga uno o veinte litros de sopa. Una muestra de tamaño $n = 100$ de una población de $N = 100mil$ elementos, tiene casi la misma precisión que una muestra de tamaño $n = 100$ de una población de $N = 100millones$ de elementos:
>
> 1. Para el primer caso, $Var_{MAS}(\hat{\bar{y}}_{\pi}) = \frac{99900}{100000} \frac{S^2_{yU}}{100} = 0.999 \dfrac{S^2_{yU}}{100}$
> 2. Para el último caso, $Var_{MAS}(\hat{\bar{y}}_{\pi}) = \frac{9999900}{100000000} \frac{S^2_{yU}}{100} = 0.999999 \dfrac{S^2_{yU}}{100}$

Tamaño de muestra

Bajo muestreo aleatorio simple sin reemplazo, un intervalo de confianza de $100(1-\alpha)\,\%$ para la media de la población es:

$$\left[\bar{y}_S - z_{1-\alpha/2}\sqrt{\left(1-\frac{n}{N}\right)}\frac{S_{yU}}{\sqrt{n}}, \bar{y}_S + z_{1-\alpha/2}\sqrt{\left(1-\frac{n}{N}\right)}\frac{S_{yU}}{\sqrt{n}}\right] \tag{3.2.14}$$

y como usualmente no se conoce $S^2_{y_U}$, lo usual es sustituirlo por el valor muestral $S^2_{y_s}$. Por lo general, sólo los investigadores del estudio pueden decidir sobre la precisión mínima del mismo. Ésta se expresa como:

$$Pr(|\bar{y}_S - \bar{y}_U| \leq c) = 1-\alpha$$

Por tanto, la cantidad a minimizar es c,

$$c = z_{1-\alpha/2}\sqrt{\left(1-\frac{n}{N}\right)}\frac{S_{yU}}{\sqrt{n}} \tag{3.2.15}$$

y despejando n, se tiene:

$$n \geq \frac{n_0}{1+\frac{n_0}{N}} \tag{3.2.16}$$

con $n_0 = \frac{z^2_{1-\alpha/2}S^2_{yU}}{c^2}$. La desigualdad se tiene porque cuando se aumenta el tamaño de muestra, c decrece su valor. En algunas ocasiones se quiere lograr una precisión relativa dada por:

$$P\left(\left|\frac{\bar{y}_S - \bar{y}_U}{\bar{y}_U}\right| \leq c\right) = 1-\alpha$$

que se puede escribir equivalentemente como:

$$P\left(|\bar{y}_S - \bar{y}_U| \leq c|\bar{y}_U|\right) = 1-\alpha$$

la cantidad a minimizar es:

$$c|\bar{y}_U| = z_{1-\alpha/2}\sqrt{\left(1-\frac{n}{N}\right)}\frac{S_{yU}}{\sqrt{n}} \tag{3.2.17}$$

y despejando n, se tiene:

$$n \geq \frac{k_0}{1+\frac{k_0}{N}} \tag{3.2.18}$$

con $k_0 = \frac{z^2_{1-\alpha/2}S^2_{yU}}{\bar{y}^2_U c^2} = \frac{z^2_{1-\alpha/2}CV^2}{c^2}$. La desigualdad se tiene porque cuando se aumenta el tamaño de muestra, $c|\bar{y}_U|$ decrece su valor.

Bajo un diseño aleatorio simple, un intervalo de confianza del $100(1-\alpha\,\%)$ para la media poblacional $\bar{y}_U$ puede ser escrito como

$$\bar{y}_S(1 \pm A) \tag{3.2.19}$$

Donde A está dada por

$$A = z_{1-\alpha/2}\sqrt{\left(1-\frac{n}{N}\right)}\frac{S_{ys}}{\sqrt{n}\bar{y}_S} = z_{1-\alpha/2}\sqrt{\left(1-\frac{n}{N}\right)}\frac{cv}{\sqrt{n}} \tag{3.2.20}$$

Asumiendo que $CV \doteq cv$ y que $\frac{n}{N}$ es una cantidad despreciable, podemos determinar un tamaño de muestra para mantener una precisión dada. Por tanto A se reescribe como

$$A \doteq z_{1-\alpha/2}\frac{CV}{\sqrt{n}}$$

y despejando n, tenemos que

$$n \geq z^2_{1-\alpha/2}\frac{CV^2}{A^2}$$

Con un nivel de confianza del $\alpha = 5\,\%$, asumiendo que el coeficiente de variación estimado converge al coeficiente de variación poblacional y que la fracción de muestreo es despreciable para obtener una precisión $A < 3\,\%$ si a) $CV = 0.5$, el tamaño de muestra debe ser mayor que 1067 unidades; b) $CV = 1.0$, el tamaño de muestra debe ser mayor que 4268 unidades y c) $CV = 1.5$, el tamaño de muestra debe ser mayor que 9604 unidades. Es decir, entre más dispersa sea la población, con respecto a la media, mayor debe ser el tamaño de muestra para conseguir una precisión dada.

Para poder utilizar las anteriores fórmulas es necesario contar un buen tamaño de muestra, dado que el teorema central del límite clásico (universo infinito) no es el mismo que se ha aplicado aquí. Hájek (1960) demuestra que al utilizar muestreo aleatorio simple (universo finito) y bajo ciertas condiciones de regularidad conocidas como las condiciones de Noether y si n, N, y $N-n$ son grandes, es decir la fracción muestral $f = n/N$ se aleja de 0 y de 1, entonces

$$\frac{\bar{y}_S - \bar{y}_U}{\sqrt{\left(1-\frac{n}{N}\right)\frac{S_{yU}}{\sqrt{n}}}} \longrightarrow Normal(0,1)$$

Cuando se quiere establecer un intervalo de confianza, la confiabilidad del intervalo está garantizada por el insesgamiento del estimador de Horvitz-Thompson. Para asegurar determinada precisión es necesario conocer la varianza poblacional de la característica de interés o el coeficiente de variación del estimador; en estos términos, cuando el coeficiente de variación estimado (cve) es menor del 3 % es un caso excelente; entre el 3 y el 5 % es bueno; entre el 5 y el 10 % es regular; entre el 10 y 15 % es apenas presentable; si es más del 15 % no es considerado bueno; en este caso algunas agencias de estadísticas oficiales no presentan el coeficiente de variación, aunque se conozca.

Por supuesto, algunas cantidades poblacionales necesarias para estimar el tamaño de muestra no se conocen; de hecho, si se conocieran, no habría necesidad de realizar estudio alguno, porque directamente se conocerían los parámetros poblacionales de interés. Lohr (2000) considera tres escenarios para realizar una estimación previa de los parámetros de interés:

1. Realizar una **prueba piloto**, unas cuantas entrevistas conforman la muestra piloto, seleccionada con el mismo diseño de muestreo genérico. En algunas ocasiones, este método además de servir para estimar las cantidades necesarias para establecer el tamaño de muestra, sirve para confrontar y calibrar el instrumento de medición, ya sea un cuestionario o un instrumento técnico.

2. Utilizar información a priori de estudios anteriores. No siempre el investigador que realiza un estudio por muestreo ha sido el primero en cuestionarse acerca de los objetivos de la investigación. Si esto es así, existen referencias bibliográficas disponibles, en donde se pueden hallar estimaciones de la varianza poblacional o del error estándar. Esta última medida tiende a ser más estable contra el tiempo o posición geográfica.

3. Estimar la varianza ajustando una distribución teórica a la característica de interés. Ospina (2001) afirma que este ajuste se hace con base en supuestos adecuados acerca de la estructura poblacional de la característica de interés (normal, exponencial, uniforme, etc.). La identificación de una distribución apropiada permite hacer uso de sus propiedades para obtener una estimación más realista de la varianza. Cuando el desconocimiento es absoluto, se recomienda utilizar la distribución uniforme. Wu (2003) afirma que las características de interés en poblaciones económicas son sesgadas a la derecha y tienden a ser modeladas mediante distribuciones como la Gamma o la Ji cuadrado.

3.2.4 Estimación en dominios

El primer caso concerniente a la estimación de subgrupo poblacionales es el de las sub-poblaciones llamadas dominios. En muchas investigaciones es necesario llevar a cabo estimaciones sobre la población en general, y también sobre subgrupos de ella (denominados dominios por la subcomisión en muestreo de las Naciones Unidas). La identificación de los dominios se logra una vez la información de los elementos ha sido registrada. Los dominios tienen que cumplir las siguientes características:

1. Ningún elemento de la población puede pertenecer a dos dominios.

2. Todo elemento de la población debe pertenecer a un único dominio.

3. La reunión de todos los dominios es la población del estudio.

Por ejemplo, al estimar el total de la fuerza laboral en empresas con menos de dos años de funcionamiento. Claramente la población se divide en dos dominios; el primero concerniente a las empresas con menos de dos años de funcionamiento y el segundo dado por las empresas con dos años o más de funcionamiento.

Definición 3.2.2. *Un dominio U_d es una sub-población específica o subgrupo poblacional que cumple las siguientes condiciones:*

1. *$U_d \subset U$, tal que $U = \bigcup_{d=1}^{D} U_d$*

2. *Si $k \in U_l$, entonces $k \notin U_d$ para $d \neq l$*

3. *El número de elementos en el dominio U_d es N_d y es llamado* ***tamaño absoluto*** *del dominio.*

4. *La proporción de elementos en el dominio U_d con respecto al tamaño poblacional es $P_d = \frac{N_d}{N}$ y se conoce como* ***tamaño relativo*** *del dominio.*

La estimación por dominios se caracteriza por el desconocimiento de la pertenencia de las unidades poblacionales al dominio. Es decir, para conocer cuáles unidades de la población pertenecen al dominio, es necesario realizar el proceso de medición.

Fue Hartley (1959) quien desarrolló y unificó la teoría de la estimación en dominios aplicable a cualquier diseño de muestreo. Durbin (1967) obtuvo resultados similares. Las pautas para la estimación en dominios se dan a continuación: para estimar el total de un dominio U_d, dado por

$$t_{yd} = \sum_{Ud} y_k \tag{3.2.21}$$

es necesario, en primer lugar construir una función indicadora z_{dk}, para cada elemento de la población, de la pertenencia del elemento al dominio, dada por la siguiente definición.

Definición 3.2.3. *Sea z_{dk} la función indicatriz del dominio U_d, dada por*

$$z_{dk} = \begin{cases} 1 & \textit{si } k \in U_d \\ 0 & \textit{en otro caso} \end{cases} \tag{3.2.22}$$

Ahora, al multiplicar la variable de pertenencia z_{dk} por el valor de la característica de interés y_k, se crea una nueva variable y_{dk} dada por $y_{dk} = z_{dk}y_k$, y una vez construida se pueden utilizar los principios del estimador de Horvitz-Thompson para hallar un estimador insesgado del total de la característica de interés en el dominio U_d.

Resultado 3.2.6. *El total de la variable de interés en el dominio U_d está dado por*

$$t_{yd} = \sum_{U} y_{dk}, \tag{3.2.23}$$

el tamaño del dominio U_d toma la siguiente expresión

$$N_d = \sum_{U} z_{dk}, \tag{3.2.24}$$

de tal forma que la media de la característica de interés en el dominio U_d se escribe como

$$\bar{y}_{U_d} = \frac{t_{yd}}{N_d} = \frac{\sum_U y_{dk}}{N_d} \tag{3.2.25}$$

Estimación del total en un dominio

Resultado 3.2.7. *Bajo muestreo aleatorio simple sin reemplazo, el estimador de Horvitz-Thompson para el total del dominio t_{yd}, su varianza y su varianza estimada*

están dados por

$$\hat{t}_{yd,\pi} = \frac{N}{n}\sum_S y_{dk} = \frac{N}{n}\sum_{S_d} y_k \tag{3.2.26}$$

$$Var(\hat{t}_{yd,\pi}) = \frac{N^2}{n}\left(1-\frac{n}{N}\right)S^2_{y_dU} \tag{3.2.27}$$

$$\widehat{Var}(\hat{t}_{yd,\pi}) = \frac{N^2}{n}\left(1-\frac{n}{N}\right)S^2_{y_dS} \tag{3.2.28}$$

respectivamente, donde $S_d = U_d \cap S$ *se refiere al conjunto formado por la intersección de la muestra* S*. Además,*

$$S^2_{y_dU} = \frac{1}{N-1}\left(\sum_{k\in U} y^2_{dk} - \frac{(\sum_{k\in U} y_{dk})^2}{N}\right)$$

representa la varianza poblacional de la característica de interés y

$$S^2_{y_dS} = \frac{1}{n-1}\left(\sum_{k\in S} y^2_{dk} - \frac{(\sum_{k\in S} y_{dk})^2}{n}\right)$$

la varianza muestral de los valores de la característica de interés.

Nótese que en la expresión $S^2_{y_dU}$ los valores que intervienen son los de la característica de interés si el elemento pertenece al dominio y ceros si el elemento no pertenece al dominio, lo mismo sucede con $S^2_{y_dS}$. Por tanto, las anteriores expresiones van a tomar valores grandes por la inclusión de los ceros; éste es el precio que se debe pagar por el desconocimiento de la pertenencia de los elementos a los dominios.

Estimación del tamaño absoluto de un dominio

Resultado 3.2.8. *Bajo muestreo aleatorio simple sin reemplazo, el estimador de Horvitz-Thompson para el tamaño absoluto de un dominio* N_d*, su varianza y su varianza estimada están dados por*

$$\hat{N}_{d,\pi} = \frac{N}{n}\sum_S z_{dk} = \frac{N}{n}\sum_{S_d} z_k \tag{3.2.29}$$

$$Var(\hat{N}_{d,\pi}) = \frac{N^2}{n}\left(1-\frac{n}{N}\right)S^2_{z_dU} \tag{3.2.30}$$

$$\widehat{Var}(\hat{N}_{d,\pi}) = \frac{N^2}{n}\left(1-\frac{n}{N}\right)S^2_{z_ds} \tag{3.2.31}$$

respectivamente, con $S^2_{z_dU}$ *y* $S^2_{z_ds}$ *la varianza poblacional y la varianza muestral de los valores de la característica de interés* z_{dk}*.*

Nótese que en la expresión $S^2_{z_dU}$ los valores que intervienen son unos si el elemento pertenece al dominio y ceros si el elemento no pertenece al dominio, lo mismo sucede con $S^2_{y_ds}$.

Estimación del tamaño relativo de un dominio

Resultado 3.2.9. *Bajo muestreo aleatorio simple sin reemplazo, el estimador de Horvitz-Thompson para el tamaño relativo de un dominio P_d, su varianza y su varianza estimada están dados por*

$$\hat{P}_{d,\pi} = \frac{1}{N}\sum_S \frac{N}{n} z_{dk} = \frac{1}{n}\sum_S z_{dk} = \frac{n_d}{n} \tag{3.2.32}$$

$$Var(\hat{P}_{d,\pi}) = \frac{1}{n}\left(1 - \frac{n}{N}\right) S^2_{z_d U} \tag{3.2.33}$$

$$\widehat{Var}(\hat{P}_{d,\pi}) = \frac{1}{n}\left(1 - \frac{n}{N}\right) S^2_{z_d s} \tag{3.2.34}$$

respectivamente, con $S^2_{z_d U}$ y $S^2_{z_d s}$ el estimador de la varianza de los valores de la característica de interés y_d en el universo y en la muestra.

Estimación de la media de un dominio

Resultado 3.2.10. *Bajo muestreo aleatorio simple sin reemplazo, el estimador de Horvitz-Thompson para la media de la característica de interés en un dominio $\bar{y}_{U_d}$, su varianza y su varianza estimada están dados por*

$$\hat{\bar{y}}_{U_d,\pi} = \frac{\frac{N}{n}\sum_S y_{dk}}{N_d} \tag{3.2.35}$$

$$Var(\hat{\bar{y}}_{U_d,\pi}) = \frac{1}{N_d^2}\frac{N^2}{n}\left(1 - \frac{n}{N}\right) S^2_{y_d U} \tag{3.2.36}$$

$$\widehat{Var}(\hat{\bar{y}}_{U_d,\pi}) = \frac{1}{N_d^2}\frac{N^2}{n}\left(1 - \frac{n}{N}\right) S^2_{y_d s} \tag{3.2.37}$$

Para poder utilizar el anterior estimador, es necesario conocer de antemano el valor del tamaño absoluto del dominio N_d. En la práctica, pocas veces se conoce este valor, por lo tanto un estimador alternativa y completamente intuitivo de la media de la característica de interés en un dominio es la media muestral de la misma en el dominio de interés. De tal forma que el estimador alternativo, toma la siguiente expresión

$$\widehat{y}_{S_d} = \frac{\hat{t}_{yd,\pi}}{\hat{N}_{d,\pi}} = \frac{\sum_S y_{dk}}{z_{dk}} = \frac{\sum_{S_d} y_k}{n_d} \tag{3.2.38}$$

Como las dos cantidades en el numerador y denominador son aleatorias, se está estimando una razón, de tal manera que el cálculo y estimación de la varianza del anterior estimador están fuera del alcance de este capítulo, y serán explicados en los lugares donde sea conveniente.

3.2.5 El efecto de diseño

Una medida que compara la eficiencia entre dos estrategias de muestreo es el efecto de diseño. Ésta herramienta práctica muestra la ganancia o pérdida, de precisión, al utilizar una estrategia de muestreo más compleja que un diseño aleatorio simple sin reemplazo junto con el estimador de Horvitz-Thompson y está definida de la siguiente manera:

Definición 3.2.4. *Siendo $(\hat{T}, p(\cdot))$ y $(\hat{T}_\pi, MAS)$ dos estrategias de muestreo utilizadas para la estimación del parámetro T, se define el efecto de diseño como*

$$Deff = \frac{Var_p(\hat{T})}{Var_{MAS}\hat{T}_\pi}. \tag{3.2.39}$$

en particular, el efecto de diseño, restringido a la estimación de un total poblacional y al usar el estimador de Horvitz-Thompson en ambas estrategias, toma la siguiente forma

$$Deff = \frac{Var_p(\hat{t}_{y,\pi})}{\frac{N^2}{n}\left(1 - \frac{n}{N}\right) S^2_{yU}}. \tag{3.2.40}$$

Cuando el efecto de diseño es más grande que la unidad, la varianza de la estrategia del numerador es más grande que la denominador, por tanto, se ha perdido precisión al utilizar una estrategia de muestreo más compleja; si el cociente es menor que uno, se ha ganado precisión. Fue Cornfield (1951) quien sugirió evaluar la eficiencia de una estrategia de muestreo al hacer el cociente entre la varianza de la misma y la del diseño aleatorio simple sin reemplazo con el estimador de Horvitz-Thompson. Más adelante Kish (1965) lo llamo DEFF (efecto de diseño, por sus siglas en inglés).

Sin embargo, en la mayoría de ocasiones, el cálculo de este cociente no es sencillo. Lehtonen & Pahkinen (2003) plantea una estimación del efecto de diseño para totales mediante la estimación de las varianzas que intervienen en la expresión. De esta forma, se tiene

Resultado 3.2.11. *Un estimador del efecto de diseño $Deff$ para el total poblacional t_y es*

$$\hat{Deff} = \frac{\widehat{Var}_p(\hat{T})}{\frac{N^2}{n}\left(1 - \frac{n}{N}\right) S^2_{ys}}. \tag{3.2.41}$$

No todos los parámetros tienen el mismo comportamiento, por lo tanto, los efectos de diseño para estos no tendrán un mismo criterio de optimalidad. Es decir, si existe un criterio de optimalidad con respecto a un parámetro, digamos el total poblacional t_y, no necesariamente se cumplirá ese criterio con un parámetro distinto, digamos la mediana poblacional.

Dado que el tamaño de muestra en diseños diferentes al muestreo aleatorio simple sin reemplazo puede ser variable, es necesario asegurarse que $n = E_{MAS}(n(S)) = E_p(n(S))$ para que exista un punto objetivo de comparación. Por ejemplo, para comparar la eficiencia del estimador de Horvitz-Thompson en el diseño de muestreo Bernoulli, es necesario fijar el tamaño de muestra, dado que este diseño no es de

tamaño fijo; es decir que $n = E_{MAS}(n(S)) = E_{BER}(n(S)) = N\pi$. Por lo que resulta que $\pi = n/N$.

De esta manera podemos introducir la medida de eficiencia del diseño de muestreo Bernoulli con respecto al MAS, así

$$deff = \frac{Var_{BER}(\hat{t}_{y,\pi})}{Var_{MAS}(\hat{t}_{y,\pi})} = 1 - \frac{1}{N} + \frac{1}{CV_y^2} \cong 1 + \frac{1}{CV_y^2} \qquad (3.2.42)$$

Por tanto, si el efecto de diseño $deff$ es igual a 1.8, esto implica que la varianza del π estimador bajo diseño de muestreo Bernoulli es 1.8 veces la varianza del π estimador bajo MAS.

3.2.6 Marco y Lucy

Una de las razones por las que el gobierno realiza la encuesta de crecimiento económico del sector industrial es, no sólo para medir el impacto social e impositivo sino para buscar nuevas estrategias de crecimiento enfocadas en las empresas que conforman este sector. Recientemente, con el boom de la tecnología y el uso masivo de internet, las estrategias de mercadeo han cambiado su forma y su fondo.

Hace unos años, las empresas con un rendimiento muy alto, catalogadas dentro de un nivel industrial grande, podían acceder a pautar un comercial discreto de 900 TRP's[2] en televisión, mientras que las empresas medianas tenían un presupuesto con el cual apenas podían pautar un comercial en la radio. Por supuesto, la estrategia publicitaria de las empresas pequeñas consistía en editar un aviso en las páginas amarillas.

Sin embargo, a medida que cambia y evoluciona la tecnología, también lo hacen los hábitos de las personas. Es muy común que las operaciones financieras, contables y estratégicas de una empresa estén centradas en un servidor conectado a internet. La misma comunicación verbal ha sido reemplazada por altos estándares de tecnología mediante conversaciones virtuales, la comunicación oficial ha desplazado el casillero de correo postal por el correo electrónico que permite la recepción en tiempo real de mensajes sin importar la ubicación espacio temporal del receptor ni de la persona que envía el mensaje. Siendo así, las personas pasan más tiempo frente a un computador que frente al televisor, o escuchando la radio; las páginas amarillas están siendo reemplazadas por los meta-buscadores de la red mundial de información, gigantes como Google, Yahoo y MSN.

Los gerentes de mercadeo (en los casos pertinentes) junto con los presidentes o gerentes de las empresas del sector industrial, han replanteado sus viejas estrategias publicitarias y han hecho, poco a poco, la migración de canal publicitario. Las empresas grandes siguen pautando en televisión, las empresas medianas siguen haciéndolo en la radio y las pequeñas siguen teniendo el mismo viejo aviso clasificado en la sección de las páginas amarillas. Sin embargo, en todos los niveles del sector industrial, se ha empezado a realizar una mejor gestión de sus clientes y/o de sus potenciales clientes.

[2]Puntos acumulados de rating del grupo objetivo obtenidos considerando sólo consumidores viendo el comercial de televisión de una marca dada

Las empresas están utilizando listas de correo electrónico masivas para dar a conocer las ventajas competitivas de sus empresas, mediante el envió de portafolios virtuales de los productos y servicios que brindan. Se cree que esta práctica de mercadeo ha aumentado la productividad empresarial porque por medio de la publicidad por internet o SPAM, las empresas consiguen más clientes, por lo tanto consiguen más contratos, por tanto ayudan a la disminución del desempleo y obtienen ventajas fiscales.

El gobierno quiere corroborar esta hipótesis y dependiendo de los resultados del estudio implementar un programa de capacitación gratuita a las empresas que aún no han entrado en el ámbito de la información mediante el uso masivo de la red informática internet. El presupuesto del gobierno es de unos cuantos millones de dólares, por lo tanto se necesitan estimaciones muy precisas que respondan al objetivo de la investigación.

Estimación del tamaño de muestra

La estrategia de muestreo que se va a utilizar es la siguiente: el estimador de Horvitz-Thompson aplicado a un diseño de muestreo aleatorio simple sin reemplazo. Se selecciona una muestra piloto de tamaño 30 de la población. Para esto, una vez cargado el archivo de datos Lucy, utilizamos la función `sample` para extraer la muestra piloto. Como la característica de interés es el ingreso de las empresas, tomamos los valores de la varianza y de la media como estimaciones que servirán para el cálculo del tamaño de la muestra.

```
> data(Lucy)
> attach(Lucy)
> N <- dim(Lucy)[1]
> sam <- sample(N,30)

> Ingresopiloto <- Income[sam]
> var(Ingresopiloto)
[1] 66952.62
> mean(Ingresopiloto)
[1] 455
```

Los valores que se utilizarán en la estimación del tamaño de muestra son la varianza muestral igual a 66.952, el promedio muestral igual a 455; con estos valores se tiene una estimación del coeficiente de variación igual a 0,57. Se debe escoger un tamaño de muestra que proporcione estimaciones precisas, el tamaño de muestra depende de la precisión que se requiera para cumplir con los objetivos del estudio.

- Error absoluto: el margen de error para este estudio es de 25 millones de dólares.
- Nivel de confianza del 95 %.
- Mediante (3.2.16) se tiene que $n_0 = 411$.
- Al utilizar el factor de corrección de poblaciones finitas, llegamos a que $n \geq 351$.

Sin embargo, este cálculo se puede cotejar restringiendo las estimaciones mediante un error relativo.

- Error relativo: se requieren estimaciones con menos del 7 % de error.
- Nivel de confianza del 95 % y una estimación de $CV = 0.57$.
- Mediante (3.2.18) se tiene que $k_0 = 446$.
- Al utilizar el factor de corrección de poblaciones finitas, llegamos a que $n \geq 376$.

Suponga que mediante fuentes oficiales se ha tenido acceso a información de estudios pasados que han modelado la característica de interesé `Income` utilizando la familia de distribuciones Gamma con parámetro de forma 2,7 y parámetro de escala 180. Haciendo una simulación de $N = 2396$ valores provenientes de una distribución gamma con los anteriores parámetros, se pueden estimar los valores de la varianza para la característica de interés y así una estimación del tamaño de muestra.

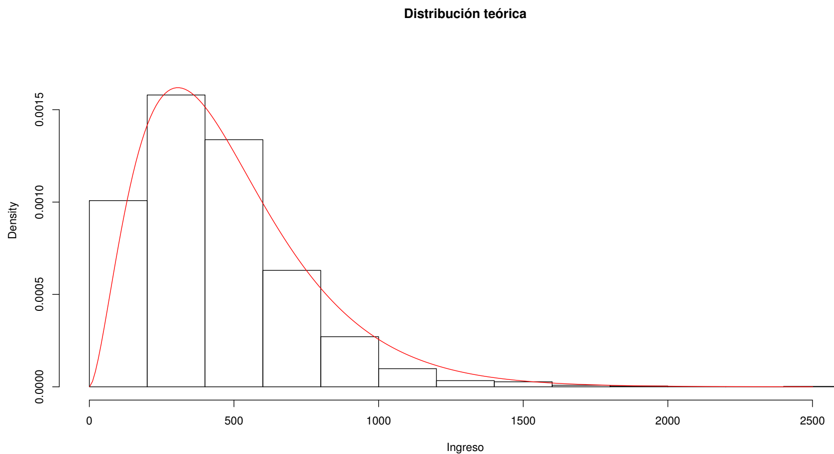


Figura 3.2: *Distribución teórica de la característica de interés: Ingreso.*

```
> teo <- rgamma(N,shape=2.7,scale=180)
> mean(teo)
[1] 484.4221
> var(teo)
[1] 92292.3
> sd(teo)/mean(teo)
[1] 0.6271318
```

La determinación del tamaño de muestra para esta investigación utilizando la estrategia de muestreo mencionada al principio de la sección y consideraciones respecto a que la estimación de la varianza de la muestra piloto puede ser pequeña, da como resultado una muestra de tamaño $n = 400$ empresas del sector industrial.

Como el tamaño de la población es $N = 2396$, entonces el valor de la probabilidad de inclusión para todos los elementos es de $\pi_k = \dfrac{400}{2396} \cong 0.17$.

R incorpora la función `sample` para la selección de muestras con o sin reemplazo. En este caso puede ser utilizada como en la selección de la muestra piloto. Sin embargo, para seleccionar una muestra mediante el algoritmo de selección y rechazo, el paquete `TeachingSampling` adjunta la función `S.SI` que se utilizará en la selección de 400 empresas del sector industrial.

Primero se carga en R el archivo `Marco` que contiene el marco de muestreo para la selección de la muestra. Se fijan los parámetros de la función, `N` y `pik`. Esta función devuelve un vector conteniendo el índice de los elementos seleccionados en la muestra. En este caso particular, el primer elemento seleccionado es el número 7 y el último el número 2395.

```
> data(Lucy)
> attach(Lucy)
> N <- dim(Lucy)[1]
> n <- 400
> sam <- S.SI(N,n)
> muestra <- Lucy[sam,]
> attach(muestra)
> muestra
          ID Ubication  Level Zone
7      AB007      c1k7  Small    A
18     AB018     c1k18  Small    A
21     AB021     c1k21  Small    A
...
2385  AB912      c26k9    Big    E
2395  AB987     c26k19    Big    E
> n <- dim(muestra)[1]
> n
[1] 400
```

Aplicando los índices obtenidos por la función `S.SI` al marco de muestreo obtenemos la identificación y ubicación de las empresas seleccionadas en la muestra. Una vez que la etapa de recolección de datos se haya realizado; es decir, la medición de todos y cada uno de los elementos seleccionados ya ha sido realizada, se realiza la estimación. Obtendremos un archivo de datos de `Lucy` conteniendo los valores de las características de interés para las empresas seleccionadas que será adjuntado a R mediante la función `attach`.

La etapa de estimación de resultados se hace utilizando la función `E.SI(N,n,y)` del paquete `TeachingSampling` cuyos argumentos son `y`, un vector conteniendo los valores de la característica de interés en la muestra, `N` el tamaño de la población y `n` el tamaño de la muestra seleccionada. En este caso la longitud de cada vector es de $n = 400$. Esta función arroja la estimación del total poblacional de `y` usando el estimador de Horvitz-Thompson, la estimación de la varianza y el coeficiente de variación del mismo. Por ejemplo, la variable `Income` dentro del objeto `estima` contiene los valores del ingreso declarado en el último año por 400 empresas del sector industrial pertenecientes a la muestra. La estimación para esta característica

se hace mediante el siguiente código:

```
> estima <- data.frame(Income, Employees, Taxes)

> E.SI(N,n,estima)
                                 Income    Employees     Taxes
Total estimado          1.009531e+06 1.497680e+05 2.607747e+04
Varianza                7.345608e+08 1.206152e+07 2.079223e+06
coeficiente de variación 2.684691e+00 2.318900e+00 5.529491e+00
```

Tabla 3.2: *Muestreo aleatorio simple: estimación de los totales de las características de interés.*

Variable	Total poblacional	Total estimado	cve %	Desv. %
Ingreso	1035217	1009531	2.68	-2.48
Empleados	151950	149768	2.31	-1.44
Impuestos	28654	26077	5.52	-8.99

La tabla 3.2. muestra los resultados obtenidos para este caso particular. Nótese que se obtienen mejores resultados que al utilizar un diseño de muestreo Bernoulli. Sin embargo, comparar estos resultados de ingreso total en el sector industrial con el de las mediciones pasadas, no es suficiente y se desea tener estimaciones para el dominio o subgrupo de las empresas que utilizan el envío de SPAM como estrategia publicitaria.

La función `Domains` contenida en el paquete `TeachingSampling` es utilizada para obtener las variables indicadoras z_{dk} para cada dominio, el único argumento de la función es un vector de pertenencia de cada individuo. En este caso, el vector de pertenencia es SPAM, la salida de esta función es una matriz de unos y ceros, en donde cada columna está dicotomizada. Existen tantas columnas como subgrupos poblacionales, y en cada columna el número uno implica la pertenencia del elemento al dominio y cero la no pertenencia del elemento al dominio.

```
> Dominios <- Domains(SPAM)
> Dominios
       no yes
  [1,]  0   1
  [2,]  0   1
  [3,]  1   0
  ...
[400,]  0   1
```

Para estimar el tamaño absoluto de cada dominio, lo único que se debe hacer es multiplicar la matriz de características de interés (en este caso, la matriz llamada `estima`) por cada columna de la matriz resultante de la dicotomización. La siguiente salida lo muestra claramente para el dominio de la población que sí utiliza el SPAM como método publicitario.

```
> SPAM.si <- Dominios[,2]*estima
> SPAM.si
    Income Employees Taxes
1      490        22  10.5
2      342        40   5.0
3        0         0   0.0
...
400   1640       225 169.0
```

Mientras que para el dominio que no utiliza el SPAM se tiene la siguiente salida

```
> SPAM.no <- Dominios[,1]*estima
> SPAM.no
    Income Employees Taxes
1        0         0   0.0
2        0         0   0.0
3      381        94   6.0
...
400      0         0   0.0
```

Utilizando la función `E.SI` en la matriz resultante de la dicotomización obtenemos las estimación de los tamaños absolutos de cada dominio. En este caso, se estima que 1420 empresas ya están utilizando otras técnicas radicales de publicidad, mientras que las restantes 976 no lo hacen. Nótese que la varianza de cada estimación es la misma, esto es claro porque los valores de esta característica de interés son ceros y uno y por tanto la estructura de varianza resulta idéntica en cada caso.

```
> E.SI(N,n,Dominios)
                                   no          si
Total estimado             976.370000 1419.630000
Varianza                  2893.945998 2893.945998
coeficiente de variación     5.509736    3.789396
```

Está claro que existe una tendencia en el sector industrial de publicidad virtual mediante el envío de SPAM por correo electrónico. Las siguientes cifras son las verdaderamente importantes pues muestran que las empresas que utilizan SPAM tienen mayores ingresos, emplean a más gente y contribuyen con una mayor cantidad de dinero en cuanto a impuestos se refiere, esto se da porque hay más empresas que utilizan el SPAM de las que no lo hacen.

```
> E.SI(N,n,SPAM.no)
                               Income    Employees        Taxes
Total estimado           4.322564e+05 6.301480e+04 1.140197e+04
Varianza                 8.666341e+08 1.719039e+07 1.293514e+06
coeficiente de variación 6.810461e+00 6.579612e+00 9.974837e+00
```

```
> E.SI(N,n,SPAM.si)
                                 Income    Employees        Taxes
Total estimado            5.772743e+05 8.675317e+04 1.467550e+04
Varianza                  9.098948e+08 1.769865e+07 1.484429e+06
coeficiente de variación 5.225326e+00 4.849363e+00 8.302079e+00
```

Como N_d es desconocido, podemos utilizar el estimador alternativo dado por la expresión (3.2.38), para obtener una estimación (aunque no la varianza ni el c.v.e) de la media de la característica de interés en cada dominio. Simplemente tomamos las estimaciones t_{yd} y las dividimos por la estimación de N_d. Las siguientes tablas resumen las estimaciones para cada uno de los dominios de interés[3].

Tabla 3.3: *Estimación en el dominio SPAM SI.*

Variable	Total poblacional	Total estimado	cve %	Desv. %
Nd	1495	1420	3.79	-5,02
Ingreso	637265	577274	5.22	-9.41
Empleados	92350	86753	4.84	-6.06
Impuestos	18061	14675	8.30	-18.75

Tabla 3.4: *Estimación en el dominio SPAM NO.*

Variable	Total poblacional	Total estimado	cve %	Desv. %
Nd	937	976	5.51	4.16
Ingreso	397752	432256	6.81	8.62
Empleados	59600	63014	6.58	5.73
Impuestos	10593	11402	9.97	7.64

3.2.7 Probabilidades de inclusión en unidades de muestreo

En Särndal, Swensson & Wretman (1992) se considera una encuesta para medir los ingresos de los hogares. El marco de muestreo es una lista de individuos y una muestra de tamaño n se selecciona mediante muestreo aleatorio simple sin reemplazo, el hogar correspondiente al individuo es identificado y se procede a realizar la medición correspondiente. La probabilidad de inclusión de un hogar h compuesto por $M < n$ individuos, puede modelarse por medio de la distribución hipergeométrica, así:

[3]Nótese que el anterior procedimiento asegura la estimación de los parámetros de dominios no sólo en MAS sino para cualquier diseño de muestreo.

$$
\begin{aligned}
\pi_H &= Pr(H \in s) \\
&= 1 - Pr(H \notin s) \\
&= 1 - Pr(\text{Ninguno de los } M \text{ salió en la muestra de tamaño } n) \\
&= 1 - \frac{\binom{M}{0}\binom{N-M}{n}}{\binom{N}{n}} \\
&= 1 - \frac{(N-M)!/n!(N-M-n)!}{N!/(N-M)!n!} \\
&= 1 - \frac{(N-M)!}{N!}\frac{(N-n)!}{(N-M-n)!} \\
&= 1 - \frac{(N-n)\dots(N-n-M+1)}{N\dots(N-M+1)}
\end{aligned}
$$

Asumiendo que N y n son grandes ($f > 0$), se obtienen las siguientes aproximaciones:

- $M = 1$,

$$
\begin{aligned}
\pi_H &= 1 - \frac{N-n}{N} \\
&= 1 - \left(1 - \frac{n}{N}\right) = 1 - (1 - f)
\end{aligned}
$$

- $M = 2$,

$$
\begin{aligned}
\pi_H &= 1 - \frac{(N-n)(N-n-1)}{N(N-1)} \\
&= 1 - \left(1 - \frac{n}{N}\right)(1 - \frac{n}{N-1}) \doteq 1 - (1-f)^2
\end{aligned}
$$

- $M = 3$,

$$
\begin{aligned}
\pi_H &= 1 - \frac{(N-n)(N-n-1)(N-n-2)}{N(N-1)(N-2)} \\
&= 1 - \left(1 - \frac{n}{N}\right)(1 - \frac{n}{N-1})(1 - \frac{n}{N-2}) \doteq 1 - (1-f)^3
\end{aligned}
$$

3.3 Muestreo aleatorio simple con reemplazo

Una **muestra aleatoria simple con reemplazo**, de tamaño m de una población de N elementos es la extracción de m muestras independientes de tamaño 1, en donde cada elemento se extrae de la población con la misma probabilidad

$$
p_k = \frac{1}{N} \qquad \forall k \in U
$$

Definición 3.3.1. *Un diseño de muestreo aleatorio simple con reemplazo se define como*

$$p(s) = \begin{cases} \frac{m!}{n_1(s)!\ldots n_N(s)!}\prod_U \left(\frac{1}{N}\right)^{n_k(s)} & si \sum_U n_k(s) = m \\ 0 & en\ otro\ caso \end{cases} \tag{3.3.1}$$

Donde $n_k(s)$ es el número de veces que el elemento k-ésimo es seleccionado en la muestra realizada s.

Resultado 3.3.1. *Para este diseño de muestreo, existen $\binom{N+m-1}{m}$ posibles muestras de tamaño m; es decir*

$$\#(Q) = \binom{N+m-1}{m}$$

Resultado 3.3.2. *Dado el soporte Q, de todas las posibles muestras con reemplazo de tamaño m, se verifica que el diseño de muestreo aleatorio simple con reemplazo es tal que*

$$\sum_{s\in Q} p(s) = 1$$

Prueba. La demostración es inmediata porque este diseño de muestro es una función de densidad multinomial discreta sobre Q.

$$\begin{aligned}
\sum_{s\in Q} p(s) &= \sum_{s\in Q} \frac{m!}{n_1(s)!\ldots n_N(s)!}\prod_U \left(\frac{1}{N}\right)^{n_k(s)} \\
&= \sum_{s\in Q} \frac{m!}{n_1(s)!\ldots n_N(s)!}\left(\frac{1}{N}\right)^{n_1(s)}\cdots\left(\frac{1}{N}\right)^{n_N(s)} \\
&= \sum_{\substack{n_1(s)\ldots n_N(s) \\ \sum_U n_k(S)=m}} \frac{m!}{n_1(s)!\ldots n_N(s)!}\left(\frac{1}{N}\right)^{n_1(s)}\cdots\left(\frac{1}{N}\right)^{n_N(s)} \\
&= \underbrace{\left(\frac{1}{N}+\cdots+\frac{1}{N}\right)^m}_{N \text{ veces}} \\
&= 1
\end{aligned}$$

donde se utiliza el resultado del teorema multinomial que afirma que

$$\sum_{\substack{n_1\ldots n_N \\ \sum_U n_k=m}} \frac{m!}{n_1!\ldots n_N!}(p_1)^{n_1}\ldots(p_N)^{n_N} = \left(\sum_{k=1}^N p_k\right)^m \tag{3.3.2}$$

■

Resultado 3.3.3. *Para un diseño aleatorio simple con reemplazo, las probabilidades de inclusión de primer y segundo orden están dadas por*

$$\pi_k = 1-\left(1-\frac{1}{N}\right)^m \tag{3.3.3}$$

$$\pi_{kl} = 1-2\left(1-\frac{1}{N}\right)^m+\left(1-\frac{2}{N}\right)^m \tag{3.3.4}$$

respectivamente.

Prueba. Utilizando los resultados 2.2.9. y 2.2.10., respectivamente, se llega a la demostración. ■

Ejemplo 3.3.1. En nuestra población ejemplo el tamaño poblacional es $N = 5$. Si se quisiera seleccionar una muestra aleatoria simple con reemplazo de tamaño $m = 2$, entonces existirían $N^m = 5^2 = 25$ posibles extracciones ordenadas. Sin embargo, sólo existen $\binom{N+m-1}{m} = \binom{6}{2} = 15$ posibles muestras. Cada una de las posibles muestras que pertenecen al soporte con reemplazo tienen las siguientes probabilidades de selección.

	V1	V2	p	n1	n2	n3	n4	n5
1	Yves	Yves	0.04	2	0	0	0	0
2	Ken	Ken	0.04	0	2	0	0	0
3	Erik	Erik	0.04	0	0	2	0	0
4	Sharon	Sharon	0.04	0	0	0	2	0
5	Leslie	Leslie	0.04	0	0	0	0	2
6	Yves	Ken	0.08	1	1	0	0	0
7	Yves	Erik	0.08	1	0	1	0	0
8	Yves	Sharon	0.08	1	0	0	1	0
9	Yves	Leslie	0.08	1	0	0	0	1
10	Ken	Erik	0.08	0	1	1	0	0
11	Ken	Sharon	0.08	0	1	0	1	0
12	Ken	Leslie	0.08	0	1	0	0	1
13	Erik	Sharon	0.08	0	0	1	1	0
14	Erik	Leslie	0.08	0	0	1	0	1
15	Sharon	Leslie	0.08	0	0	0	1	1

Nótese que la suma de las probabilidades inducidas por el diseño de muestreo es igual a uno y que cada una de ellas es mayor que cero.

3.3.1 Algoritmo de selección

Tillé (2006) presenta dos algoritmos para seleccionar una muestra aleatoria simple con reemplazo. El primero, de manera general induce m selecciones individuales y el segundo, es un método secuencial que implementa la selección mediante la distribución binomial.

Método de m selecciones

El siguiente método de selección se implementa en m pasos, y aunque no es eficiente computacionalmente, es muy conocido.

- Seleccionar un primer elemento con probabilidad $\frac{1}{N}$ de todo el conjunto de datos.
- Seleccionar un segundo elemento con probabilidad $\frac{1}{N}$ de todo el conjunto de datos.

- ...
- Seleccionar un m-ésimo elemento con probabilidad $\frac{1}{N}$ de todo el conjunto de datos.

Hace unas pocas décadas, cuando no existía la ayuda tecnológica de ahora, no imagino como los encargados de la selección de la muestra pudieron haber utilizado este algoritmo. Imagine seleccionar una muestra de 3000 elementos sin la facilidad de un computador.

Método secuencial

Tillé (2006) afirma que este procedimiento es mejor que el anterior porque permite seleccionar una muestra de tamaño m en una sola pasada por el conjunto de datos.

- Seleccionar n_k veces el elemento k-ésimo de acuerdo a una distribución binomial.

  $$Bin\left(m - \sum_{i=1}^{k-1} n_i, \frac{1}{N-k+1}\right) \tag{3.3.5}$$

  Para todo $k \in U$.

Ejemplo 3.3.2. Como se ha visto en los capítulos anteriores, R incorpora en la función `sample`, la selección de muestras aleatorias simples con reemplazo, simplemente el argumento replace debe ser activado mediante, `replace=TRUE`. Así, para seleccionar una muestra con reemplazo de tamaño $m = 3$, sólo es necesario escribir el siguiente código.

```
> sam  <-  sample(N,3, replace=TRUE)
> U[sam]
[1] "Erik"   "Leslie" "Leslie"
```

El procedimiento de selección de una muestra aleatoria con reemplazo de tamaño m mediante el uso del algoritmo secuencial está implementado en la función `S.WR(N,m)` cuyos argumentos son `N`, el tamaño de la población y `m`, el tamaño de la muestra con reemplazo. Así, para seleccionar una muestra aleatoria simple con reemplazo de la población U de tamaño $N = 5$, se tiene

```
> m <- 3
> sam <- S.WR(N,m)
> U[sam]
[1] "Ken"    "Sharon" "Sharon"
```

Una vez más, la salida de la función es un vector de índices (no necesariamente distintos) de los elementos pertenecientes a la muestra seleccionada s. Este algoritmo utiliza la distribución binomial en cada uno de sus pasos, de tal forma que para la selección de la anterior muestra conformada por **Ken**, **Sharon** y **Sharon** cada uno de los $N = 5$ pasos del algoritmo arrojaron los siguientes resultados.

k	Nombre	Bin n	Bin p	nk
1	Yves	3	0.2000	0
2	Ken	3	0.2500	1
3	Erik	2	0.3333	0
4	Sharon	2	0.5000	2
5	Leslie	0	1.0000	0

Donde `Bin n` y `Bin p` son los parámetros de la distribución binomial asociada al algoritmo secuencial. Note que la cantidad `nk` se refiere a la realización de la variable $n_k(s)$.

3.3.2 El estimador de Hansen-Hurwitz

Cuando se tienen las cantidades del resultado 3.3.3 se pueden implementar los principios del estimador de Horvitz-Thompson para estimar el total poblacional t_y; sin embargo, el cálculo y estimación de la varianza de esta estrategia de muestreo resulta ser muy compleja (computacionalmente). Por esta razón, utilizaremos el estimador de Hansen-Hurwitz dado por (2.2.34) que estima de manera insesgada al parámetro de interés t_y.

Resultado 3.3.4. *Para un diseño de muestreo aleatorio simple con reemplazo, el estimador de Hansen-Hurwitz del total poblacional t_y, su varianza y su varianza estimada están dados por:*

$$\hat{t}_{y,p} = \frac{N}{m}\sum_{i=1}^{m} y_i \tag{3.3.6}$$

$$Var_{MRAS}(\hat{t}_{y,p}) = N\frac{(N-1)}{m}S^2_{yU} \tag{3.3.7}$$

$$\widehat{Var}_{MRAS}(\hat{t}_{y,p}) = \frac{N^2}{m}S^2_{ysr} \tag{3.3.8}$$

respectivamente, con S^2_{yU} el estimador de la varianza de los valores de la característica de interés y en el universo y S^2_{ysr} el estimador de la varianza de los valores y_i que pertenecen a la muestra seleccionada ($\forall i \in m$) (no necesariamente distintos) en la muestra. Esto es,

$$S^2_{ysr} = \frac{1}{m-1}\sum_{i=1}^{m}(y_i - \bar{y}_S)^2.$$

Nótese que $\hat{t}_{y,p}$ es insesgado para el total poblacional t_y de la característica de interés y, y que $\widehat{Var}_{MRAS}(\hat{t}_{y,p})$ es insesgado para $Var_{MRAS}(\hat{t}_{y,p})$.

Prueba. Los resultados se obtienen escribiendo el estimador de Hansen-Hurwitz de la siguiente manera,

$$\hat{t}_{y,p} = \frac{1}{m}\sum_U n_k(S)\frac{y_k}{p_k} = \frac{N}{m}\sum_U n_k(S)y_k \tag{3.3.9}$$

Por tanto, utilizando el resultado 2.2.8., se tiene que

$$\begin{aligned} E\left(\hat{t}_{y,p}\right) &= \frac{N}{m}\sum_U E(n_k(S))y_k \\ &= \frac{N}{m}\sum_U \frac{m}{N}y_k = t_y \end{aligned}$$

Por otro lado, asumiendo que las variables Z_i son independientes e idénticamente distribuidas

$$\begin{aligned} Var\left(\hat{t}_{y,p}\right) &= Var\left(\frac{1}{m}\sum_i^m Z_i\right) \\ &= \frac{1}{m^2}\sum_i^m Var(Z_i) \\ &= \frac{1}{m^2}\sum_i^m\left(\sum_U \frac{1}{N}(Ny_k - t)^2\right) \\ &= \frac{1}{m}\left(\frac{N^2}{N}\sum_U (y_k - \bar{y}_U)^2\right) \\ &= N\frac{(N-1)}{m}S^2_{yU} \end{aligned}$$

Escribiendo el estimador de la varianza como

$$\widehat{Var}(\hat{t}_{y,p}) = \frac{1}{m}\frac{1}{m-1}\sum_U n_k(S)\left(Ny_k - \hat{t}_{y,p}\right)^2 \tag{3.3.10}$$

se tiene el insesgamiento dado por

$$\begin{aligned} E\left(\widehat{Var}(\hat{t}_{y,p})\right) &= \frac{1}{m}\frac{1}{m-1}\sum_U E\left(n_k(S)(Ny_k - \hat{t}_{y,p})^2\right) \\ &= \frac{1}{m}\frac{1}{m-1}\sum_U E\left(n_k(S)(Ny_k - t_y)^2 - n_k(S)(\hat{t}_{y,p} - t_y)^2\right) \\ &= \frac{1}{m}\frac{1}{m-1}E\left(\sum_U n_k(S)(Ny_k - t_y)^2\right) \\ &\qquad - \frac{1}{m}\frac{1}{m-1}E\left((\hat{t}_{y,p} - t_y)^2\sum_U n_k(S)\right) \\ &= \frac{1}{m}\frac{1}{m-1}\left[E\left(\sum_U n_k(S)(Ny_k - t_y)^2\right) - mE\left((\hat{t}_{y,p} - t_y)^2\right)\right] \\ &= \frac{1}{m}\frac{1}{m-1}\left[m\left(\sum_U \frac{m}{N}(Ny_k - t_y)^2\right) - mVar(\hat{t}_{y,p})\right] \\ &= \frac{1}{m}\frac{1}{m-1}\left[m^2Var(\hat{t}_{y,p}) - mVar(\hat{t}_{y,p})\right] \\ &= Var(\hat{t}_{y,p}) \end{aligned}$$

■

Ejemplo 3.3.3. Para nuestra población de ejemplo U, existen $\binom{N+m-1}{m} = 20$ posibles muestras con reemplazo de tamaño $m = 2$. Realice el cálculo léxico-gráfico del estimador de Hansen-Hurwitz y compruebe el insesgamiento y la varianza.

3.3.3 Marco y Lucy

Suponga que se quiere seleccionar una muestra aleatoria simple con reemplazo de tamaño $m = 400$ empresas del sector industrial. Para la selección de la muestra es posible usar la función `sample` que viene integrada con R. En primer lugar se debe cargar el marco de muestreo que permite la selección, identificación y posterior ubicación de cada individuo en la muestra con reemplazo. Para la selección de la muestra es necesario ingresar los parámetros de la función, en este caso `N`=2396, el tamaño poblacional, está dado por la cantidad de filas (registros de empresas del sector industrial) del marco de muestro y `m`=400 empresas que se seleccionaran con reemplazo.

```
> data(Lucy)
> attach(Lucy)
> N <- dim(Lucy)[1]
> m <- 400
> sam <- sample(N,m, replace=TRUE)
```

Sin embargo, para seleccionar la muestra con reemplazo utilizando el método secuencial, el paquete `TeachingSampling` adjunta la función `S.WR` cuyos argumentos son `N`, el tamaño de la población y `m`, el tamaño de la muestra con reemplazo. El resultado de la función es un conjunto de índices (no necesariamente distintos) que aplicados a la población resulta en los valores de la característica de interés para las empresas (no necesariamente distintas) seleccionadas. Nótese que una empresa seleccionada se tendrá en cuenta en la etapa de estimación tantas veces como haya sido seleccionada.

```
> sam <- S.WR(N,m)
> muestra <- Lucy[sam,]
> attach(muestra)
> muestra
           ID Ubication  Level Zone
2       AB002      c1k2  Small    A
15      AB015     c1k15  Small    A
15.1    AB015     c1k15  Small    A
...
2383    AB906     c26k7    Big    E
2383.1  AB908     c26k8    Big    E
2392    AB984    c26k16    Big    E
> dim(muestra)
[1] 400   4
```

La primera empresa en ser seleccionada mediante el método secuencial es la empresa que ocupa la segunda posición en el marco de muestreo; es decir, la empresa cuyo número único de identificación corresponde a **AB002**, la segunda y tercera

empresa en ser seleccionadas corresponde a la empresa identificada con el número único **AB015**. Si un elemento ha sido seleccionada más de una vez, R codifica automáticamente las posteriores selecciones con un punto seguido de un número que indica el número de veces menos uno que ha sido seleccionada la misma unidad.

Una vez que las empresas son seleccionadas, se programa la visita del encuestador en la cual se registran los valores de las características de interés. Cuando se tiene la base de datos con la información pertinente para todas las empresas seleccionadas en la muestra con reemplazo, se procede a estimar los totales de las características de interés. La función `E.WR` del paquete `TeachingSampling` permite la estimación de una o varias características de interés simultáneamente. Para ello, se debe crear un conjunto de datos con la información recolectora para cada una de las 400 empresas en las características de interés. En este caso creamos un conjunto de datos con las tres características de interés `Income`, `Employees` y `Taxes`.

```
> estima <- data.frame(Income, Employees, Taxes)
```

La función `E.WR` del paquete `TeachingSampling` tiene tres argumentos, `N`, el tamaño de la población y `m`, el tamaño de la muestra con reemplazo y el conjunto de datos (conteniendo los valores para la(s) característica(s) de interés). El resultado de la función es la estimación del total, la varianza estimada y el respectivo coeficiente de variación de la(s) característica(s) de interés.

```
> E.WR(N,m,estima)
                               Income    Employees        Taxes
Total estimado           1.049137e+06 1.569140e+05 3.082753e+04
Varianza                 1.255743e+09 1.847564e+07 6.356860e+06
coeficiente de variación 3.377679e+00 2.739290e+00 8.178667e+00
```

La tabla 3.5. muestra los resultados particulares de esta estrategia de muestreo. Nótese que con un menor tamaño de muestra, se obtienen mejores resultados que al utilizar una estrategia de muestreo que contempla un diseño Bernoulli y el estimador de Horvitz-Thompson.

Tabla 3.5: *Muestreo aleatorio simple con reemplazo: estimación de los totales de las características de interés.*

Variable	Total poblacional	Total estimado	cve %	Desv. %
Ingreso	1035217	1049137	3.37	1.34
Empleados	151950	156914	2.73	3.27
Impuestos	28654	30817	8.17	7.59

El efecto de diseño

Sin embargo, utilizando el efecto de diseño podemos comparar la eficiencia de la anterior estrategia utilizada en Lucy mediante el efecto de diseño. Utilizando la definición podemos aproximar la medida mediante

$$
\begin{aligned}
Deff &= \frac{Var_{MRAS}(\hat{t}_{y,p})}{Var_{MAS}(\hat{t}_{y,\pi})} \\
&= \frac{1}{1-f}\left(1-\frac{1}{N}\right) \cong \frac{1}{1-f}
\end{aligned}
$$

Por tanto, para la estrategia de muestreo utilizada anteriormente, tenemos $Deff = 1.2$. Lo anterior indica que existe una pérdida del 20 % de precisión al utilizar la estrategia de muestreo con reemplazo y el estimador de Hansen-Hurwitz. En general se tiene que para tamaños de muestra muy pequeños, en comparación a N, las dos estrategias arrojan resultados muy similares. Sin embargo, a medida que el tamaño de muestra crece, en comparación a N, la medida $Deff$ aumenta significativamente; es decir, existe una pérdida muy grande de eficiencia.

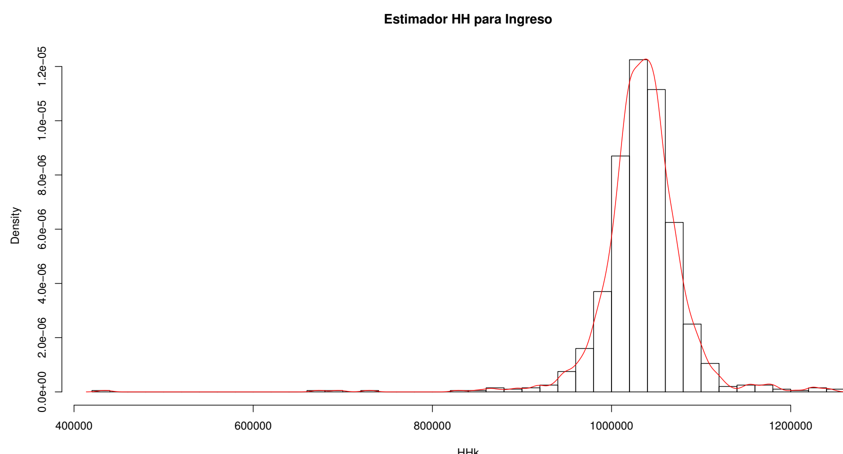


Figura 3.3: *Distribución del estimador de Hansen-Hurwitz para el total de Ingreso.*

Dado que el diseño de muestreo es con reemplazo, se quiere verificar que la distribución asintótica del estimador de Hansen-Hurwitz sea normal. Se realiza una simulación de Monte Carlo, con los mismos lineamentos utilizados en la sección 3.1.3 en donde se realizaron varios experimentos de Monte Carlo para examinar el comportamiento del estimador de Hansen-Hurwitz en la característica ingreso. El resultado de la simulación se muestra en los histogramas de la figura 3.3. En este experimento de Monte Carlo el promedio de las estimaciones de cada experimento coincide con el total poblacional y se espera que la varianza de las estimaciones debe acercarse a la varianza basada en el diseño de muestreo aleatorio simple.

La media de las estimaciones de t_y es 1034179 que ajusta bien con el parámetro correspondiente $t_y = 1035217$. Nótese que la varianza del estimador (mediante este experimento de Monte Carlo) es muy grande y que la distribución del estimador muestra valores atípicos, aunque con una probabilidad de ocurrencia muy baja. Hay que tener cuidado con las afirmaciones acerca de normalidad en este caso pues la distribución, aunque parece ser simétrica y con forma de campana, está sesgada a derecha.

3.4 Diseño de muestreo sistemático

En algunas ocasiones, cuando no se dispone de un marco de muestreo, por lo menos no de forma explícita, o cuando el marco disponible está ordenado de forma particular, con respecto a los rótulos del mismo, es posible utilizar el diseño de muestreo sistemático como una opción para la selección de muestras. La característica más particular de este diseño de muestreo es que todas las unidades se suponen enumeradas del 1 al N, al menos implícitamente, y se tiene conocimiento de que la población se encuentra particionada en a grupos poblacionales latentes. En este orden de ideas el tamaño poblacional N puede ser escrito como

$$N = na + c \tag{3.4.1}$$

en donde $0 \leq c < a$ y n, el tamaño de muestra esperado, se define como la parte entera del cociente N/a. Nótese que c es un entero que representa el residuo algebraico del total poblacional y se puede ver fácilmente que toma la siguiente forma

$$c = N - \left\| \frac{N}{a} \right\| a \tag{3.4.2}$$

En donde $\| \frac{N}{a} \|$ representa la parte entera del cociente N/a. Una vez que los grupos han sido conformados, se procede a escoger de manera aleatoria, un número entre 1 y a, por ejemplo r. La muestra estará conformada sistemáticamente por los elementos $r, r+a, r+2a, \ldots, r+(n-1)a$. Nótese que en el caso en donde $c = 0$, el tamaño de muestra estará dado por $n = N/a$; de otra forma, si $c > 0$, el tamaño de muestra puede ser $n = \| \frac{N}{a} \|$ ó $n = \| \frac{N}{a} \| + 1$. Como lo señala Raj (1968) este diseño de muestreo es un caso especial de un muestreo por conglomerados, como se verá en los siguientes capítulos.

Tabla 3.6: *Posible configuración del muestreo sistemático.*

Grupo	s_1	$\cdots$	s_r	$\cdots$	s_a
$n = 1$	1	$\cdots$	r	$\cdots$	a
$n = 2$	$1 + a$	$\cdots$	$r + a$	$\cdots$	$2a$
$n = 3$	$1 + 2a$	$\cdots$	$r + 2a$	$\cdots$	$3a$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
$n = \| \frac{N}{a} \|$	$1 + (n-1)a$	$\cdots$	$r + (n-1)a$	$\cdots$	na
$n = \| \frac{N}{a} \| + 1$	$1 + na$	$\cdots$	$\square$	$\cdots$	$\square$

El anterior esquema permite una mejor comprensión del funcionamiento del diseño de muestreo sistemático. Nótese el ordenamiento por grupos de las unidades que pertenecen a la población. En particular, esta tabla corresponde a una población, en donde, si se seleccionara el último grupo s_a, entonces el tamaño de muestra sería $n = \| \frac{N}{a} \|$, mientras que si se escogiera el primer grupo s_1, el tamaño de muestra estaría dado por $n = \| \frac{N}{a} \| + 1$.

Por otro lado, nótese que cada grupo s_r constituye una posible muestra, de tal forma que

$$U = \bigcup_{r=1}^{a} s_r. \tag{3.4.3}$$

El soporte Q de todas las posible muestras sistemáticas, queda entonces definido como

$$Q_r = \{s_1, s_2, \ldots, s_r, \ldots, s_a\}. \tag{3.4.4}$$

Resultado 3.4.1. *Para este diseño de muestreo, la cardinalidad del soporte es igual al número de grupos formados. Es decir*

$$\#Q_r = a$$

Definición 3.4.1. *Suponga que el tamaño poblacional es tal que $N = na + c$, con $0 \leq c < a$. Se define un diseño de muestreo sistemático de la siguiente manera*

$$p(s) = \begin{cases} \frac{1}{a} & \text{si } s \in Q_r \\ 0 & \text{en otro caso} \end{cases} \tag{3.4.5}$$

Dado que sólo existen a posibles muestras, el diseño de muestreo sistemático cumple que $\sum_{s \ni Q} p(s) = 1$.

3.4.1 Algoritmo de selección

El siguiente algoritmo secuencial permite la extracción de una muestra mediante el diseño de muestreo sistemático.

1. Seleccionar con probabilidad $\dfrac{1}{a}$ un arranque aleatorio. Es decir un entero r, tal que $1 \leq r \leq a$.

2. La muestra estará definida por el siguiente conjunto

$$s_r = \{k : k = r + (j-1)a; j = 1, \ldots, n(S)\} \tag{3.4.6}$$

Ejemplo 3.4.1. Nuestra población ejemplo U está ordenada de la siguiente forma

$$U = \{\textbf{Yves, Ken, Erik, Sharon, Leslie}\}$$

Suponga que sistemáticamente se divide en $a = 2$ grupos. El primero dado por:

$$s_1 = \{\textbf{Yves, Erik, Leslie.}\}$$

y el segundo conformado por:

$$s_2 = \{\textbf{Ken, Sharon}\}$$

De tal forma que $N = (2)(2) + 1$. Para seleccionar un arranque aleatorio r se utilizará un dado, de tal forma que si el resultado de un lanzamiento es par, entonces la muestra seleccionada será s_1, de lo contrario la muestra seleccionada será s_2.

Resultado 3.4.2. *Para un diseño de muestreo sistemático, las probabilidades de inclusión de primer y segundo orden están dadas por*

$$\pi_k = \frac{1}{a} \tag{3.4.7}$$

$$\pi_{kl} = \begin{cases} \frac{1}{a} & \text{si } k \text{ y } l \text{ pertenecen a } s_r \\ 0 & \text{en otro caso} \end{cases} \tag{3.4.8}$$

respectivamente.

Prueba. considerando que el elemento k-ésimo sólo puede pertenecer a una y sólo una muestra s_r, tenemos que

$$\pi_k = Pr(k \in S) = Pr(\text{seleccionar la muestra } s_r) = \frac{1}{a} \tag{3.4.9}$$

Por otra parte, suponga que los elementos k-ésimo y l-ésimo pertenecen al grupo s_r. De esta manera, estos elementos son incluidos en la muestra sí y sólo sí se selecciona el grupo s_r, por tanto, la probabilidad de inclusión de segundo orden está dada por la probabilidad de selección del grupo s_r igual a $\frac{1}{a}$. Si los elementos k-ésimo y l-ésimo pertenecen a grupos distintos, la probabilidad de ser incluidos en la muestra realizada es nula. ■

3.4.2 El estimador de Horvitz-Thompson

Una vez que el diseño de muestreo es definido, la estrategia se completa con el uso del estimador de Horvitz-Thompson, por ser este un diseño sin reemplazo. El siguiente resultado será útil para definir las propiedades de varianza del estimador.

Resultado 3.4.3. *Para un diseño $p(\cdot)$ con soporte Q, la varianza del estimador de Horvitz-Thompson, se puede escribir como*

$$Var(\hat{t}_{y,\pi}) = \sum\sum_U \frac{\pi_{kl}}{\pi_k\pi_l} y_k y_l - \left(\sum_U y_k\right)^2 \tag{3.4.10}$$

Prueba. Partiendo del resultado 2.2.2., se tiene que

$$\begin{aligned} Var(\hat{t}_{y,\pi}) &= \sum\sum_U \Delta_{kl} \frac{y_k}{\pi_k}\frac{y_l}{\pi_l} && (3.4.11) \\ &= \sum\sum_U (\pi_{kl} - \pi_k\pi_l) \frac{y_k}{\pi_k}\frac{y_l}{\pi_l} && (3.4.12) \\ &= \sum\sum_U \left(\frac{\pi_{kl}}{\pi_k\pi_l} - 1\right) y_k y_l && (3.4.13) \\ &= \sum\sum_U \frac{\pi_{kl}}{\pi_k\pi_l} y_k y_l - \sum\sum_U y_k y_l && (3.4.14) \\ &= \sum\sum_U \frac{\pi_{kl}}{\pi_k\pi_l} y_k y_l - \left(\sum_U y_k\right)^2 && (3.4.15) \end{aligned}$$

En donde se utiliza el hecho de que

$$\sum\sum_{U} y_k y_l = \sum\sum_{k \neq l} y_k y_l + \sum_{U} y_k^2 = \left(\sum_{U} y_k\right)^2 \tag{3.4.16}$$

■

Resultado 3.4.4. *Para el diseño de muestreo sistemático, el estimador de Horvitz-Thompson y su varianza están dados por:*

$$\hat{t}_{y,\pi} = at_{sr}, \tag{3.4.17}$$

con $t_{sr} = \sum_{k \in S_r} y_k$, *y*

$$Var_{SIS}(\hat{t}_{y,\pi}) = a \sum_{r=1}^{a} (t_{sr} - t)^2 \tag{3.4.18}$$

En este caso no existe estimador de la varianza.

Prueba. De la definición del estimador de Horvitz-Thompson y dado que las probabilidades de inclusión de primer orden son todas iguales al valor $1/a$, entonces

$$\hat{t}_{y,\pi} = \sum_{Sr} \frac{y_k}{\pi_k} = at_{sr} \tag{3.4.19}$$

Utilizando los dos anteriores resultados, se sigue que

$$Var(\hat{t}_{y,\pi}) = \sum\sum_{U} \frac{\pi_{kl}}{\pi_k \pi_l} y_k y_l - \left(\sum_{U} y_k\right)^2 \tag{3.4.20}$$

$$= a \sum_{r=1}^{a} \left(\sum\sum_{sr} y_k y_l\right) - t^2 \tag{3.4.21}$$

$$= a \sum_{r=1}^{a} \left(\sum_{k \in s_r} y_k \sum_{l \in sr} y_l\right) - t^2 \tag{3.4.22}$$

$$= a \sum_{r=1}^{a} t_{s_r}^2 - t^2 \tag{3.4.23}$$

$$= a \sum_{r=1}^{a} (t_{sr} - \bar{t})^2 \tag{3.4.24}$$

donde

$$\bar{t} = \sum_{r=1}^{a} \frac{t_{s_r}}{a} = \frac{t}{a} \tag{3.4.25}$$

Por la definición 3.4.1, algunas probabilidades de inclusión de segundo orden son nulas, por ell no se tiene un estimador de la varianza. ■

Más allá de que los principios del estimador de Horvitz-Thompson no permitan estimar la varianza para este diseño, la razón genérica radica en que, de una forma u otra, se está seleccionando uno y sólo un grupo de elementos y se calcula un sólo total para el grupo. Como la selección es de sólo un grupo, no se tiene un marco de comparación y no se puede llegar a una estimación de la varianza.

3.4.3 Optimalidad de la estrategia

Una vez que la estrategia de muestreo queda definida, es indispensable tocar el tema de la configuración de los valores de la característica de interés mediante el ordenamiento particular que se tiene en el marco de muestreo. Bautista (1998) utiliza el siguiente esquema para explicar la eficiencia de esta estrategia de muestreo.

Tabla 3.7: *Configuración de totales por grupo.*

Grupo	s_1	$\cdots$	s_r	$\cdots$	s_a
	y_1		y_r		y_k
Valor de	y_{1+a}		y_{r+a}		y_{2a}
la	y_{1+2a}		y_{r+2a}		y_{3a}
característica	$\cdots$		$\cdots$		$\cdots$
	$y_{1+(n-1)a}$		$y_{r+(n-1)a}$		y_{na}
Total de grupo	t_{s_1}	$\cdots$	t_{s_r}	$\cdots$	t_{s_a}

Este diseño de muestreo puede resultar más eficiente que el diseño de muestreo aleatorio simple, dependiendo del ordenamiento del marco de muestreo. Es usado para palear las posibles imperfecciones generadas por un diseño de muestreo aleatorio simple. Por ejemplo, puede resultar que en una muestra simple, todos los elementos de la muestra seleccionada compartan una característica latente que perjudique la precisión de las estimaciones. En el caso de una población de personas, puede resultar que una muestra simple sólo incluya hombres. Cuando se sabe que el marco de muestreo está ordenado de manera aleatoria, es recomendable utilizar el diseño de muestreo aleatorio simple, porque asegura una muestra bien mezclada. Por ejemplo, si el marco de muestreo está ordenado alfabéticamente, es casi seguro que se obtendrá una muestra que sea representativa de la población, puesto que la posición alfabética no debería estar asociada con la característica de interés.

Además, mediante este diseño de muestreo, no es necesario poseer un marco de muestreo de forma física para poder realizar una muestra probabilística. Sin embargo, se debe tener cuidado con la especificación del diseño, pues como lo afirma Lohr (2000) no es lo mismo seleccionar una de cada 10 personas que entran a una biblioteca que seleccionar una de cada 10 personas que salen de un avión. En el segundo caso, existe de forma implícita, un marco de muestreo.

Como se verá más adelante, el diseño de muestreo sistemático puede ser más preciso que el diseño de muestreo aleatorio simple cuando los grupos s_r poseen mucha variación interna. De manera contraria, si el valor de los elementos dentro de los grupos proporciona la misma información, entonces la eficiencia del diseño se verá disminuida significativamente con respecto al diseño aleatorio simple.

La figura 3.4 muestra los tres casos más particulares en el uso de esta estrategia de muestreo cuyas características son las siguientes:

1. **Ordenamiento aleatorio**: cuando el ordenamiento del marco de muestreo no está relacionado con la característica de interés, la eficiencia de este diseño es comparable con la de muestreo aleatorio simple. Ordenamiento por orden alfabético.

2. **Ordenamiento lineal**: cuando el ordenamiento del marco de muestreo es tal que se puede observar una tendencia lineal, entonces la selección de una muestra sistemática obliga a que los valores de los elementos incluidos tengan una alta dispersión haciendo que el comportamiento de los grupos formados sea heterogéneo con respecto al valor de la característica de interés. Ordenamiento de registros contables.

3. **Ordenamiento periódico**: si la población es tal que se observa un patrón de tipo periódico, el muestreo sistemático puede arrojar peores resultado que una muestra aleatoria simple pues si el intervalo de muestreo coincide con el patrón de periodicidad, la muestra seleccionada incluiría elementos cuyos valores de la característica de interés serían muy parecidos. Una muestra seleccionada de esta manera no sería representativa de la población. En algunos casos es posible encontrar poblaciones con este tipo de comportamiento periódico; por ejemplo, el flujo vehicular durante las 24 horas del día o las ventas en negocios durante cierta temporada del año.

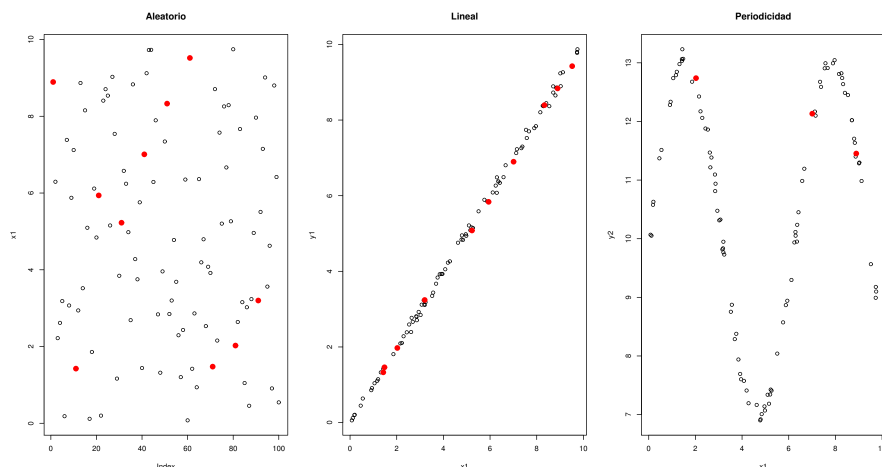


Figura 3.4: *Casos de ordenamiento en muestreo sistemático.*

Descomposición de la varianza

Algunos críticos de la teoría del muestreo han querido separar el pensamiento estadístico de la metodología de estudios por muestreo. Lo anterior sumado a la falta de preparación del usuario del muestreo ha abierto una brecha entre dos mundos. La verdad es que la estadística sin muestreo no está completa y viceversa Kish (1965). En estos apartes, debemos considerar uno de los resultados más importantes de la estadística que ha permitido el desarrollo de la misma en diversos campos de la vida práctica.

Resultado 3.4.5. *Suponga que la población se divide en a grupos, de tal forma que existen n elementos por grupo y el tamaño poblacional toma la forma $N = an$, entonces*

$$(N-1)S^2_{y_U} = \underbrace{\sum_U (y_k - \bar{y}_U)^2}_{SCT} = \underbrace{\sum_{r=1}^{a}\sum_{s_r}(y_{rk} - \bar{y}_{s_r})^2}_{SCD} + \underbrace{\sum_{r=1}^{a} n\,(\bar{y}_{s_r} - \bar{y}_U)^2}_{SCE} \quad (3.4.26)$$

La sigla **SCT** se refiere a la suma de cuadros del total de la población y no es otra cosa que el numerador en la fórmula del estimador de la varianza. El anterior resultado es importante porque permite descomponer la suma de cuadrados total en dos cantidades. Primero, **SCD** que denota la suma de cuadrados dentro (al interior) de los grupos y segundo, **SCE** que hace referencia a la suma de cuadrados entre los grupos. Por supuesto, la varianza como parámetro poblacional es fija, por tanto si

1. **SCE** es alta, entonces **SCD** es baja, indicando así que los grupos están construidos de tal forma que resultan ser muy heterogéneos entre sí, pero dentro de ellos existe homogeneidad.
2. **SCE** es baja, entonces **SCD** es alta, lo que quiere decir que los grupos son muy disímiles en su interior, pero entre ellos tienen un comportamiento similar.

Esta representación de la descomposición de la varianza, se puede ver claramente en una tabla de ANOVA (análisis de varianza, por sus siglas en inglés), de la siguiente manera.

Tabla 3.8: *Tabla de ANOVA inducida por el muestreo sistemático.*

Fuente	gl	Suma de cuadrados	Cuadrado medio
Entre	$a-1$	$SCE = \sum_{r=1}^{a} n\,(\bar{y}_{s_r} - \bar{y}_U)^2$	$\dfrac{SCE}{a-1}$
Dentro	$N-a$	$SCD = \sum_{r=1}^{a}\sum_{s_r}(y_{rk} - \bar{y}_{s_r})^2$	$\dfrac{SCD}{N-a}$
Total	$N-1$	$SCT = \sum_U (y_k - \bar{y}_U)^2$	$s^2_{y_U}$

Desde un punto de vista totalmente pragmático, la estrategia de muestreo tendrá un mejor desempeño cuando la variabilidad total entre los grupos sea mínima y la variabilidad dentro de los grupos sea máxima. El siguiente resultado da una mejor comprensión de la descomposición de la varianza en los grupos. Es decir, la varianza del estimador de Horvitz-Thompson, bajo muestreo sistemático, será cercana a cero cuando el ordenamiento de los grupos en la población es tal que los totales t_{s_r} con $r = 1, \ldots, a$ son similares

$$t_{s_1} \approx t_{s_2} \approx \cdots \approx t_{s_a} \approx \bar{t} \quad (3.4.27)$$

Resultado 3.4.6. *Sin pérdida de generalidad, considere que el tamaño muestral es tal que $N = na$, entonces la varianza del estimador de Horvitz-Thompson bajo*

un diseño de muestreo sistemático toma la siguiente forma

$$Var_{SIS}(\hat{t}_{y,\pi}) = N\sum_{r=1}^{a} n\left(\bar{y}_{s_r} - \bar{y}_U\right)^2 = N(SCE) \tag{3.4.28}$$

Prueba. Partiendo de la definición de la varianza del estimador de Horvitz-Thompson en muestreo sistemático, se tiene que

$$\begin{aligned}
Var_{SIS}(\hat{t}_{y,\pi}) &= a\sum_{r=1}^{a}\left(t_{sr} - \bar{t}\right)^2 \\
&= \frac{N}{n}\sum_{r=1}^{a}\left(n\bar{y}_{sr} - n\bar{y}_U\right)^2 \\
&= \frac{N}{n}\sum_{r=1}^{a} n^2\left(\bar{y}_{sr} - \bar{y}_U\right)^2 \\
&= N\sum_{r=1}^{a} n\left(\bar{y}_{s_r} - \bar{y}_U\right)^2 = N(SCE)
\end{aligned}$$

■

Por tanto, se quiere que toda la variabilidad esté por dentro de cada uno de los grupos.

Definición 3.4.2. *Se define el coeficiente de correlación intra-clase como*

$$\rho = 1 - \frac{n}{n-1}\frac{SCD}{SCT} \tag{3.4.29}$$

Esta medida de correlación entre los pares de elementos de los grupos formados toma una valor máximo igual a uno cuando **SCE** es nula y toma un valor mínimo igual a $-\frac{1}{n-1}$ cuando **SCE** es máxima. En particular, es deseable para esta estrategia que ρ tome valores cercanos a cero.

Resultado 3.4.7. *Utilizando la relación 3.4.26* ***SCT=SCE+SCD*** *se tiene que*

$$SCE = SCT\left[(\rho - 1)\frac{n-1}{n} + 1\right] \tag{3.4.30}$$

Prueba. De la definición del coeficiente de correlación intra-clase se tiene que

$$\begin{aligned}
(\rho - 1)\frac{n-1}{n} + 1 &= 1 - \frac{SCD}{SCT} \\
&= \frac{SCE}{SCT}
\end{aligned}$$

por tanto al despejar SCE se tiene el resultado. ■

Resultado 3.4.8. *Con el anterior resultado no es difícil verificar que la varianza del estimador de Horvitz-Thompson bajo muestreo sistemático se puede escribir como*

$$Var_{SIS}(\hat{t}_{y,\pi}) = \underbrace{\frac{N^2}{n}\left(1 - \frac{n}{N}\right)S^2_{yU}}_{Var_{MAS}(\hat{t}_{y,\pi})}\left\{\frac{N-1}{N-n}[1 + (n-1)\rho]\right\} \tag{3.4.31}$$

Prueba. Partiendo de la última expresión tenemos que

$$\begin{aligned}\frac{N^2}{n}\left(1-\frac{n}{N}\right)S^2_{yU}\left\{\frac{N-1}{N-n}[1+(n-1)\rho]\right\}&=\frac{N}{n}SCT\,[1+(n-1)\rho]\\&=N(SCT)\left[1-\frac{SCD}{SCT}\right]\\&=N(SCE)\\&=Var_{SIS}(\hat{t}_{y,\pi})\end{aligned}$$

que coincide con la varianza del estimador de Horvitz-Thompson en muestreo sistemático ■

Nótese que la primera parte de la anterior ecuación se refiere al valor del estimador de Horvitz-Thompson bajo un diseño de muestreo aleatorio simple sin reemplazo. Siguiendo esta idea, el efecto de diseño está dado por el siguiente resultado.

Resultado 3.4.9. *El efecto de diseño de la estrategia de muestreo que utiliza un diseño sistemático y el estimador de Horvitz-Thompson está dado por*

$$Deff=\frac{Var_{SIS}\hat{t}_\pi}{Var_{MAS}\hat{t}_\pi}=\frac{N-1}{N-n}[1+(n-1)\rho] \tag{3.4.32}$$

Dado el efecto de diseño, se concluye que esta estrategia de muestreo es

1. *Igual de eficiente al muestreo aleatorio simple sí* $\rho=\frac{1}{1-N}$.
2. *Menos eficiente que el muestreo aleatorio simple sí* $\rho>\frac{1}{1-N}$.
3. *Más eficiente que el muestreo aleatorio simple sí* $\rho<\frac{1}{1-N}$.

Prueba. La demostración es inmediata teniendo en cuenta el anterior resultado. ■

3.4.4 Diseño de muestreo q-sistemático

Cuando la periodicidad es un problema o cuando se quiere tener un estimativo insesgado de la varianza del estimador de Horvitz-Thompson, Mahalanobis (1946) propone el uso de muestras sistemáticas inter-penetradas. Este método consiste en seleccionar, no una, sino q muestras sistemáticas. De esta manera se seleccionan q arranques aleatorios en grupos de tamaño aq, de tal manera que el tamaño poblacional se escribe como $N=a\frac{n}{q}+c$.

Definición 3.4.3. *El diseño de muestreo sistemático con q réplicas está definido como*

$$p(s)=\frac{1}{\binom{a}{q}} \qquad \textit{para todo } s\in Q_r \tag{3.4.33}$$

con Q_r definido en 3.4.4.

Por supuesto, la cardinalidad del soporte es $\#Q_r = \binom{a}{q}$, por tanto este diseño de muestreo cumple las propiedades del capítulo anterior. Teniendo en cuenta que se han formado a grupos, entonces el diseño de muestreo q-sistemático puede ser visto como un diseño MAS de tamaño de muestra igual a q de los totales de todos los grupos. Una ve más, estos grupos también pueden ser vistos como conglomerados.

Resultado 3.4.10. *Para un diseño de muestreo sistemático, las probabilidades de inclusión de primer y segundo orden están dadas por*

$$\pi_k = \frac{q}{a} \tag{3.4.34}$$

$$\pi_{kl} = \begin{cases} \frac{q}{a} & \text{si } k \text{ y } l \text{ pertenecen a} s_r \\ \frac{q}{a}\frac{q-1}{a-1} & \text{en otro caso} \end{cases} \tag{3.4.35}$$

respectivamente.

Resultado 3.4.11. *Para el diseño de muestreo sistemático con q réplicas, el estimador de Horvitz-Thompson y su varianza están dados por:*

$$\hat{t}_{y,\pi} = \frac{a}{q}\sum_S t_{sr} \tag{3.4.36}$$

$$Var_{SIS}(\hat{t}_{y,\pi}) = \frac{a^2}{q}\left(1-\frac{q}{a}\right)S^2_{t_{s_r}U} \tag{3.4.37}$$

$$\widehat{Var}_{SIS}(\hat{t}_{y,\pi}) = \frac{a^2}{q}\left(1-\frac{q}{a}\right)S^2_{t_{s_r}s} \tag{3.4.38}$$

respectivamente, con $S^2_{t_{s_r}U}$ y $S^2_{t_{s_r}s}$ el estimador de la varianza de los totales de la característica de interés y en cada grupo s_r del universo y en la muestra. Nótese que $\hat{t}_{y,\pi}$ es insesgado para el total poblacional t_y de la característica de interés y, y que $\widehat{Var}_{SIS}(\hat{t}_{y,\pi})$ es insesgado para $Var_{SIS}(\hat{t}_{y,\pi})$.

Al respecto de esta estrategia, el lector debe notar que:

- La varianza del estimador de Horvitz-Thompson bajo el diseño de muestro q-sistemático crece cuando se aplica a un universo que está ordenado igualmente de forma sistemática.
- La varianza del estimador de Horvitz-Thompson bajo el diseño de muestro q-sistemático depende del ordenamiento de los valores de la característica de interés por lo que puede suceder que ésta no sea monótonamente decreciente en función del tamaño de muestra.
- El efecto de la correlación intra-clase tiene una gran repercusión en el tamaño de muestra; si existe una alta correlación intra-clase entonces el tamaño de muestra debe ser mayor para tener un *c.v.e* pequeño y viceversa.
- En estudios de tipo electoral se dice que un candidato tiene alta correlación intra-clase (por ejemplo en los barrios) cuando la imagen del candidato está polarizada. Es decir, la mayoría de votación en determinado barrio es muy alta por el candidato o muy baja. Por otro lado, se dice que la campaña electoral tiene baja correlación intra-clase cuando la votación en los barrios no es ni muy baja ni muy alta.

3.4.5 Marco y Lucy

En nuestro intento de obtener estimaciones precisas para la evaluación del comportamiento del sector industrial en lo corrido del último año fiscal, hemos notado que el marco de muestreo está ordenado de manera alfanumérica en orden ascendente por el rótulo de identificación industrial. Además, se sabe que el número de identificación de cada empresa no tiene una secuencia específica, sino que es asignado de acuerdo a la fecha de registro de la empresa. De tal forma, la primera empresa en ser registrada ante el organismo gubernamental competente es la identificada con el número de identificación **AB001** y la última empresa en ser registrada es la identificada con el número **AB987**.

Nótese que las característica de interés son Ingreso, número de empleados e impuestos declarados en el último año fiscal y se supone, de manera correcta, que estas características no tienen ninguna relación con la fecha de registro de la empresa. Así, puede suceder que una empresa joven, tenga unos altos réditos, pocos empleados y una alta declaración de impuestos, pero puede suceder lo contrario; de hecho, este comportamiento está sujeto a la estrategia de *marketing* utilizada en cada periodo comercial y no a la antigüedad del negocio. Por las anteriores razones, se supone que el ordenamiento del marco de muestreo es completamente aleatorio.

Se ha decidido que la población va a ser particionada en seis grupos, de tal forma que el tamaño efectivo de muestra será 399 o 400. El marco de muestreo es cargado en el ambiente de R.

```
> data(Lucy)
> attach(Lucy)
> N <- dim(Lucy)[1]
> a <- 6
> floor(N/a)
> 399
```

El procedimiento que se sigue es la creación de los grupos sistemáticos. Esto puede realizarse con la función `(array(1:a,N))` que permite la creación de la secuencia **1,2,3,4,5,6,1,2,3,4,5,6,1,2...**; sin embargo, es indispensable definir este arreglo como un factor, es decir como una variable de tipo categórica nominal cuyos rótulos significan la pertenencia de un individuo a un grupo. Así el marco de muestreo toma la siguiente forma.

```
> grupo <- as.factor(array(1:a,N))
> data.frame(grupo, Lucy)[1:12,]
  grupo    ID Ubication Level Zone
1     1 AB001      c1k1 Small    A
2     2 AB002      c1k2 Small    A
3     3 AB003      c1k3 Small    A
4     4 AB004      c1k4 Small    A
5     5 AB005      c1k5 Small    A
6     6 AB006      c1k6 Small    A
7     1 AB007      c1k7 Small    A
8     2 AB008      c1k8 Small    A
```

```
9       3 AB009      c1k9 Small    A
10      4 AB010     c1k10 Small    A
11      5 AB011     c1k11 Small    A
12      6 AB012     c1k12 Small    A
...
```

La selección de la muestra se realiza mediante la función `S.SY` del paquete `TeachingSampling` cuyos argumentos son `N`, el tamaño de la población y `a`, el número de grupos. Esta función sigue el algoritmo secuencial descrito en esta estrategia de muestreo y lo que hace es aleatoriamente asignar un arranque aleatorio y saltar, en este caso, de seis en seis elementos hasta barrer toda la lista. El resultado de la función es un listado de índices que aplicados a la población resulta en los valores de las características de interés de los elementos incluidos en la muestra realizada.

```
> sam <- S.SY(N,a)
> muestra <- Lucy[sam,]
> attach(muestra)
> muestra
         ID Ubication  Level Zone
3     AB003      c1k3  Small    A
9     AB009      c1k9  Small    A
15    AB015     c1k15  Small    A
...
2379  AB892     c26k3    Big    E
2385  AB912     c26k9    Big    E
2391  AB983    c26k15    Big    E
> n <- dim(muestra)[1]
> n
[1] 399
```

En el anterior caso particular, el arranque aleatorio fue igual a tres; por tanto, la muestra está conformada por los elementos **3, 9, ..., 2385 y 2391** del marco de muestreo. Una vez recolectada la información de la muestra, se procede a realizar la estimación mediante el uso de la función[4] `E.SY` del paquete `TeachingSampling` cuyos argumentos son `N`, `a` y un conjunto de datos conteniendo la información de las características de interés para cada elemento en la muestra.

```
> estima <- data.frame(Income, Employees, Taxes)
> E.SY(N, a, estima)
                                Income    Employees        Taxes
Total estimado            1.054295e+06 1.524350e+05 3.058000e+04
Varianza                  7.652192e+08 1.196375e+07 3.658156e+06
coeficiente de variación  2.623800e+00 2.269075e+00 6.254515e+00
```

Los resultados de la estimación se muestran en la tabla 3.9. Es de considerar que la eficiencia de esta estrategia de muestreo es mucho mayor a la de una estrategia

[4]Dado que no existe el estimador genérico para la varianza del estimador de Horvitz-Thompson, esta función utiliza una aproximación conservadora de la varianza suponiendo que se realizó un muestreo aleatorio simple.

que utilice un diseño de muestreo aleatorio simple. Nótese que los coeficientes de variación son mucho menores y también, aunque este es un argumento un poco más débil, la desviación relativa es menor.

Tabla 3.9: *Muestreo sistemático: estimación de los totales de las características de interés.*

Variable	Total poblacional	Total estimado	cve %	Desv. %
Ingreso	1035217	1054295	2.62	1.84
Empleados	151950	152435	2.27	0.32
Impuestos	28654	30580	6.25	6.72

Es hora de preguntarse, ¿por qué los resultados de las estimaciones son mejores que en otro tipo de estrategias de muestreo? Vamos a realizar un procedimiento de evaluación, puramente académico, y vamos a suponer que tenemos acceso a la información de la característica de interés a nivel poblacional.

En primer lugar, se realiza un análisis de varianza para obtener la descomposición de las sumas de cuadrados para la característica de interés `Income`. Para esto usamos la función `lm` que relaciona a la variable de interés con un factor de agrupamiento. La variable grupo fue creada como un vector de cinco niveles y puede ser usada en este caso. Aplicando la función `anova` al modelo, se obtiene una tabla de sumas de cuadrados.

```
> data(Lucy)
> attach(Lucy)
> anova(lm(Income~grupo))
Analysis of Variance Table

Response: Income
            Df    Sum Sq Mean Sq F value Pr(>F)
grupo        5     12359    2472  0.0346 0.9994
Residuals 2390 170698187   71422
```

Siguiendo a Dalgaard (2008), en la mayoría de textos estadísticos (incluyendo el que el lector tiene en sus manos) las sumas de cuadrados son rotuladas como `SCD`, `SCE` y `SCT`. Sin embargo, R usa una rotulación diferente. La variación **entre** los grupos es rotulada con el nombre del factor de agrupación, en este caso `grupo`. La variación **dentro** de los factores de agrupación es rotulada como `Residuals`. Por tanto, se observa que la variación total se encuentra dentro de los grupos; mientras que existe una baja variación entre los grupos. Esto es bueno para efectos de la eficiencia de la estrategia.

Por un lado, al observar la gráfica de la característica de interés con respecto al ordenamiento natural del marco de muestreo, no es posible identificar un patrón lineal o de periodicidad, cuando realizamos el gráfico con respecto a los grupos, nos damos cuenta de que dentro de ellos existe una muy alta variabilidad y más aún, los cinco grupos tiene un comportamiento parecido entre ellos. El código necesario para la creación de este gráfico está dado a continuación.

```
> par(mfrow=c(1,2))
```

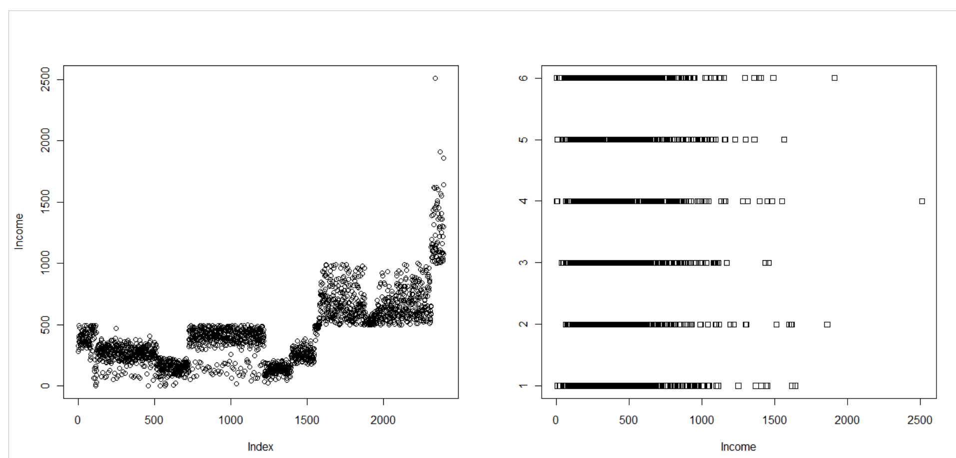


Figura 3.5: *Nube de puntos de Ingreso y su respectiva variación de los grupos.*

```
> plot(Income)
> stripchart(Income~grupo)
```

Por otro lado, el ordenamiento aleatorio se observa muy claramente en la figura 3.6., en dónde los puntos marcados corresponden a los elementos seleccionados. Nótese la buena dispersión de la muestra en la población, haciéndola representativa. El código necesario para la creación de este gráfico es el siguiente.

```
> plot(Income)
> points(sam, Income[sam],col = "red", pch=19)
```

Es claro que esta estrategia de muestreo resulto más eficiente que la estrategia de muestreo aleatorio simple. Pero, ¿cuánto más eficiente?. Con unos simple cálculos algebraicos se obtiene un coeficiente de correlación intra-clase muy cercano a cero y esto es bueno puesto que cumple con los requerimientos en la definición de ρ.

```
> rho <- 1-(n/(n-1))*(170698187/(170698187+12359))
> rho
[1] -0.002439984
> rho < 1/(1-N)
[1] TRUE
```

Sin embargo, lo verdaderamente asombroso es que la ganancia en eficiencia al usar este diseño es de veintinueve veces puesto que el efecto de diseño es aproximadamente 0.034.

```
> Deff <- (N-1)*(1+(n-1)*rho)/(N-n)
> Deff
[1] 0.03464363
> 1/Deff
[1] 28.86534
```

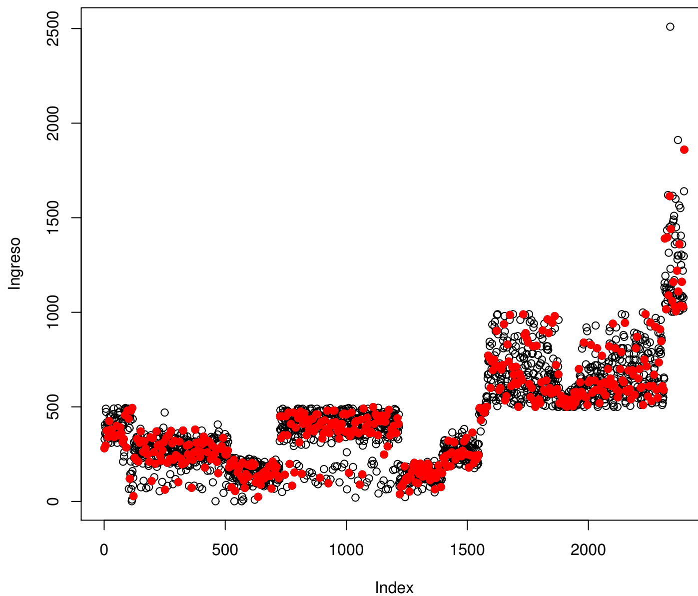


Figura 3.6: *Casos seleccionados en muestreo sistemático.*

```
> VarHT <- N*12359
> VarHT
[1] 29612164
```

Los anteriores diseños de muestreo pertenecen al grupo de los diseños de probabilidad de inclusión constante. En el siguiente capítulo veremos diseños con probabilidad de inclusión proporcional al tamaño que hace uso de información auxiliar continua en el marco de muestreo.

3.5 Ejercicios

3.1 Suponga una población de 10 elementos $U = \{e_1, e_2, \ldots, e_{10}\}$.

- Seleccione una muestra mediante un diseño Bernoulli con probabilidad de inclusión $\pi = 0.4$, utilizando el algoritmo de la sección 3.1.1. y teniendo en cuenta que para cada elemento en la población se obtuvo el siguiente conjunto de números aleatorios uniformes

$$\varepsilon = \{0.152, 0.158, 0.614, 0.593, 0.140, 0.851, 0.803, 0.996, 0.433, 0.790\}$$

- Otra manera de seleccionar una muestra Bernoulli es generando un sólo número aleatorio de una distribución $Binomial(N, \pi)$; este valor generado es el tamaño de muestra $n(S)$ y con ayuda del marco de muestreo

se selecciona una muestra aleatoria simple de tamaño $n(S)$. Suponiendo que la realización de $Binomial(10, 0.4)$ fue $n(s) = 5$, utilice el algoritmo coordinado negativo para la selección de una muestra, teniendo en cuenta que para cada elemento en la población se obtuvo el siguiente conjunto de números aleatorios uniformes

$$\boldsymbol{\xi} = \{0.370, 0.561, 0.064, 0.412, 0.952, 0.461, 0.256, 0.275, 0.213, 0.443\}$$

3.2 Complete el cálculo léxico-gráfico del ejemplo 3.1.1.

3.3 En un estudio de calidad de vida en cárceles, se utilizó un diseño de muestreo Bernoulli con probabilidad de inclusión $\pi = 0.15$ para seleccionar una muestra de reclusos. En la penitenciaría hay 1243 reclusos y se observaron las características de interés **CVDP** y **OTMA** para los presos incluidos en la muestra. Además se obtuvieron los siguientes resultados

Característica	$\sum_s y_k$	$\sum_s y_k^2$
CVDP	5412	95299
OTMA	82503	604926

- Utilice el estimador de Horvtiz-Thompson para calcular una estimación del total poblacional, el coeficiente de variación estimado y un intervalo de confianza al 95 % para estas características de interés.
- Utilice el estimador de Horvtiz-Thompson para calcular una estimación de la media poblacional, el coeficiente de variación estimado y un intervalo de confianza al 95 % para estas características de interés.
- Si el tamaño de muestra efectivo fue 191, utilice el estimador alternativo para calcular una estimación del total poblacional y de la media poblacional.

3.4 Suponga una población de 12 elementos $U = \{e_1, e_2, \ldots, e_{12}\}$. Seleccione una muestra aleatoria simple sin reemplazo de tamaño $n = 4$ utilizando el algoritmo de Fan-Muller-Rezucha teniendo en cuenta que para cada elemento en la población se obtuvo el siguiente conjunto de números aleatorios uniformes

$$\boldsymbol{\xi} = \{0.787, 0.946, 0.766, 0.338, 0.520, 0.849, 0.828, 0.165, 0.416, 0.105, 0.069, 0.853\}$$

3.5 Complete el cálculo léxico-gráfico del ejemplo 3.2.2.

3.6 Demuestre o refute la siguiente afirmación: «En muestreo aleatorio simple, para la estimación de un total poblacional, el estimador de Horvitz-Thompson coincide con el estimador altervativo».

3.7 Demuestre o refute la siguiente afirmación: «En muestreo aleatorio simple, para la estimación de un total en dominios de interés, se cumple siempre que $\sum_{d=1}^{D} \hat{t}_{yd,\pi} > \hat{t}_{y,\pi}$».

3.8 Demuestre o refute la siguiente afirmación: «En muestreo aleatorio simple, el coeficiente de variación estimado del estimador de Horvitz-Thompson para el total poblacional es menor que el coeficiente de variación estimado del estimador de Horvitz-Thompson para la media poblacional».

3.9 En un estudio de satisfacción empresarial en una entidad prestadora de salud que sirve a 748 asociados, se quiere averiguar el promedio del número de horas al mes (**NHM**) que los asociados permanecen en consulta médica. Para esto se planea un muestreo aleatorio simple pues se conoce que, para este caso particular, una aproximación para la varianza de esta característica de interés es de 3.4839 y para el coeficiente de variación es de 0.5324.

- Con una confianza del 95 %, determine el tamaño de muestra mínimo para estimar el parámetro de interés con un error absoluto no mayor 15 minutos.
- Con una confianza del 95 %, determine el tamaño de muestra mínimo para estimar el parámetro de interés con un erro relativo no mayor a 2 %.

3.10 Demuestre las siguientes igualdades

$$(n-1)S^2_{yS} = \sum_{k\in S}(y_k - \bar{y}_S)^2 = \sum_{k\in S} y_k^2 - \frac{(\sum_{k\in S} y_k)^2}{n}$$

$$(N-1)S^2_{yU} = \sum_{k\in U}(y_k - \bar{y}_U)^2 = \sum_{k\in U} y_k^2 - \frac{(\sum_{k\in U} y_k)^2}{N}$$

3.11 Demuestre rigurosamente los resultados 3.2.7 y 3.2.8.

3.12 Para el ejercicio 3.9, suponga que se deciden realizar $n = 50$ entrevistas y que se obtuvo que $\sum_s y_k = 178$ y $\sum_s y_k^2 = 826$. A continuación se presenta una tabla de frecuencias de las observaciones

NHM	0	1	2	3	4	5	6	7	8
Frecuencia	1	5	13	9	7	4	6	4	1

- Obtenga una estimación de Horvitz-Thompson para el total de horas mensuales que los asociados permanecen en consulta médica, reporte el coeficiente de variación estimado y un intervalo de confianza al 95 %.
- Obtenga una estimación de Horvitz-Thompson para el promedio de horas mensuales que los asociados permanecen en consulta médica, reporte el coeficiente de variación estimado y un intervalo de confianza al 95 %.
- Obtenga una estimación de Horvitz-Thompson para el total de asociados que permanecen en consulta médica menos (estrictamente) de cuatro horas, reporte el coeficiente de variación estimado y un intervalo de confianza al 95 %.
- Obtenga una estimación de Horvitz-Thompson para la proporción de asociados que permanecen en consulta médica, más (estrictamente) de seis horas, reporte el coeficiente de variación estimado y un intervalo de confianza al 95 %.

3.13 Complete el cálculo léxico-gráfico del ejemplo 3.3.3.

3.14 Para una población de $N = 10$ elementos se planeó diseño aleatorio simple con reemplazo de tamaño de muestra $m = 6$. Complete la siguiente salida del algoritmo secuencial utilizado para la extracción de la muestra

```
  k    nbin        pbin   nk
 [1,]                      0
 [2,]    6    0.1111111    3
 [3,]                      1
 [4,]    2    0.1428571    0
 [5,]         0.1666667    1
 [6,]    1
 [7,]    1    0.2500000    0
 [8,]                      0
 [9,]    1                 0
[10,]    1                 1
```

3.15 Suponga que se realizó un muestreo aleatorio simple con reemplazo para la población del ejercicio 3.3.

- Utilice el estimador de Hansen-Hurwitz para obtener una estimación del total poblacional para características de interés **CVDP** y **OTMA**, reporte el coeficiente de variación estimado y un intervalo de confianza del 95 %.
- Bajo el supuesto de muestreo aleatorio simple con reemplazo, construya las probabilidades de inclusión de primer y segundo orden y utilice el estimador de Horvitz-Thompson para calcular una nueva estimación del total poblacional para las características de interés.

3.16 Demuestre o refute la siguiente afirmación: «Para tamaños de muestra iguales, la estrategia de muestreo aleatorio simple con reemplazo junto con el estimador de Hansen-Hurwitz es siempre de menor varianza que la estrategia de muestreo aleatorio simple sin reemplazo junto con el estimador de Horvitz-Thompson».

3.17 Demuestre o refute la siguiente afirmación: «El diseño de muestreo sistemático es de tamaño de muestra fijo».

3.18 Demuestre o refute la siguiente afirmación: «Aunque no existe la estimación de la varianza del estimador de Horvitz-Thompson en muestreo sistemático, es siempre conveniente reemplazarla por la expresión de la varianza estimada en un diseño aleatorio simple».

3.19 Para estimar el total de horas diarias que los estudiantes permanecen en la biblioteca de una universidad, se utilizó un diseño de muestreo sistemático con dos arranques aleatorios. La población fue divida en siete grupos latentes y se seleccionó una muestra simple de dos enteros entre el uno y el siete. Los enteros seleccionados son el 3, y 7. Lo anterior implica que la muestra de estudiantes, que serán entrevistados a la salida de la biblioteca, está conformada por dos grupos. A saber el grupo s_3 conformado por los estudiantes 3, 10, 17, ... y el grupo s_7 conformado por los estudiantes 7, 14, 21, ...Los resultados del sondeo para los dos grupos se dan acontinuación

$$t_{s_3} = \sum_{s_3} y_k = 3574 \qquad t_{s_7} = \sum_{s_7} y_k = 5024$$

Calcule una estimación insesgada para el número total de horas de permanencia en la biblioteca, reporte el coeficiente de variación estimado y un intervalo de confianza al 95 %.

3.20 Suponga una población de 9 elementos cuyos valores para la característica de interés se dan a continuación

$$\mathbf{y} = \{23, 20, 24, 31, 24, 29, 25, 33, 21\}$$

- Utilice el análisis de varianza (ANOVA) para calcular la varianza del estimador de Horvitz-Thompson en un diseño de muestreo sistemático simple con $a = 2$ grupos.
- Calcule el coeficiente de variación intra-clase y el efecto de diseño. Decida si, para este caso particular, el diseño sistemático es más eficiente que el diseño de muestreo aleatorio simple.

3.21 Demuestre o refute la siguiente afirmación: «En un diseño de muestreo sistemático, si hay homogeneidad dentro de los grupos y heterogeneidad entre sus medias, entonces este diseño es menos eficiente que el diseño de muestreo aleatorio simple».

Capítulo 4

Muestreo con probabilidades proporcionales

> Es bien sabido que la estrategia de muestreo que utiliza un diseño de muestreo aleatorio simple con el estimador de Horvitz-Thompson, es una estrategia de muestreo óptima, bajo ciertas formulaciones, si se tiene un conocimiento a priori de que el comportamiento de la población es simétrico con respecto a los rótulos. En tales casos, la incorporación de información auxiliar no mejora la anterior estrategia.
>
> Claes-Magnus Cassel (1976)

Las estrategias de muestreo implementadas en el capítulo anterior, utilizaban métodos de selección tales que la probabilidad de inclusión o probabilidad de selección es idéntica para todos los elementos de la población y se estimaban los parámetros de interés utilizando el estimador de Hansen-Hurwitz, para diseños de muestreo con reemplazo y el estimador de Horvitz-Thompson, para diseños de muestreo sin reemplazo. Las anteriores estrategias no tienen en cuenta la variación innata de las características de interés a través de las unidades poblacionales. Por lo tanto, los anteriores estimadores, dada su construcción genérica y el principio de representatividad, tenderán a poseer una gran variación.

Raj (1968) afirma que, en cuestión de precisión, se puede tener una mayor ganancia cuando se utilizan diseños de muestreo con probabilidades desiguales. En la mayoría de los casos prácticos, la característica de interés no presenta un comportamiento uniforme con respecto a los rótulos de la población. Sin embargo, cuando el marco de muestreo disponible para la selección de la muestra contiene además de la identificación y la ubicación de los elementos en la población, una característica auxiliar continua disponible para todos los elementos de la población $x_k \quad \forall k \in U$, es posible utilizar diseños de muestreo que implementen métodos de selección cuyas probabilidades de selección o inclusión, dependiendo del caso, sean proporcionales al total de la característica auxiliar, t_x.

4.1 Diseño de muestreo de Poisson

Este diseño de muestreo es una generalización del diseño de muestreo Bernoulli, en donde las probabilidades de inclusión están dadas a priori de manera independiente para cada individuo. Brewer (2002) indica que este diseño de muestreo no tuvo originalmente ninguna implicación práctica, porque el tamaño de muestra no es fijo, sino que fue utilizado de manera teórica para describir las propiedades de otros estimadores. El primer caso práctico se dio en la selección de muestras de árboles en unidades forestales; más adelante se aplicó en el censo anual manufacturero en Estados Unidos. Aunque este diseño de muestreo no utiliza información auxiliar para la selección de la muestra, sirve como punto de partida para examinar diseños de muestreo más complejos que sí lo utilizan.

Definición 4.1.1. *Siendo π_k un número positivo, tal que $0 < \pi_k \leq 1$, que representa la probabilidad de inclusión del k-ésimo elemento, el diseño de muestreo Poisson se define de la siguiente manera*

$$p(s) = \prod_{k\in s} \pi_k \prod_{k\notin s}(1 - \pi_k) \quad \text{para todo } s \in Q \tag{4.1.1}$$

con Q, el soporte que contiene a todas las posibles muestras sin reemplazo.

Resultado 4.1.1. *Para este diseño de muestreo, el soporte Q tiene cardinalidad igual a*

$$\#(Q) = 2^N$$

Ejemplo 4.1.1. En nuestra población ejemplo

$$U = \{\text{Yves}, \text{Ken}, \text{Erik}, \text{Sharon}, \text{Leslie}\}$$

Las probabilidades de inclusión π_k son 0.2, 0.5, 0.7, 0.5 y 0.9, respectivamente. Las posibles muestra pueden ser de tamaño 0, 1, 2, 3, 4 ó 5. La probabilidad de la muestra de tamaño 0 es

$$(1 - 0.2) \times (1 - 0.5) \times (1 - 0.7) \times (1 - 0.5) \times (1 - 0.9) = 0.006$$

Siguiendo esta misma analogía, a continuación se presenta el cálculo léxico-gráfico para las probabilidades de selección de todas las posible muestras en el soporte de este diseño de muestreo. Para las posibles muestras de tamaño 1, 4 se tiene que sus respectivas probabilidades son:

```
  s        p(s)    |          s                   p(s)
 Yves      0.0015  |   Yves, Ken, Erik, Sharon      0.0035
 Ken       0.006   |   Yves, Erik, Sharon, Leslie   0.0315
 Erik      0.014   |   Yves, Ken, Erik, Leslie      0.0315
 Sharon    0.006   |   Yves, Ken, Sharon, Leslie    0.0135
 Leslie    0.054   |   Ken, Erik, Sharon, Leslie    0.126
 Total     0.0815  |          Total                 0.206
```

Las posibles muestras de tamaño 2, 3 y sus respectivas probabilidades son:

s	p(s)	s	p(s)
Yves, Ken	0.0015	Yves, Ken, Erik	0.0035
Yves, Erik	0.0035	Yves, Ken, Sharon	0.0015
Yves, Sharon	0.0015	Yves, Ken, Leslie	0.0135
Yves, Leslie	0.0135	Yves, Erik, Sharon	0.0035
Ken, Erik	0.014	Yves, Erik, Leslie	0.0315
Ken, Sharon	0.006	Yves, Sharon, Leslie	0.0135
Ken, Leslie	0.054	Ken, Erik, Sharon	0.014
Erik, Sharon	0.014	Ken, Erik, Leslie	0.126
Erik, Leslie	0.126	Ken, Sharon, Leslie	0.054
Sharon, Leslie	0.054	Erik, Sharon, Leslie	0.126
Total	0.288	Total	0.387

Finalmente, la muestra de tamaño 5, $\{\text{Yves}, \text{Ken}, \text{Erik}, \text{Sharon}, \text{Leslie}\}$, tiene probabilidad 0.0315. Nótese que la suma de todas las posibles muestras es $\sum p(s) = 1$.

4.1.1 Algoritmo de selección

Bautista (1998) afirma que el conocimiento a priori de las probabilidades de inclusión de los elementos es tal que, en algunas ocasiones, existen elementos de la población que deben ser observados obligatoriamente en la muestra, en estos casos el valor de la probabilidad de inclusión de estos elementos es igual a uno ($\pi_k = 1$). Al subgrupo poblacional cuyos elementos tienen probabilidad de inclusión igual a uno, se le conoce como subgrupo de **inclusión forzosa**. Nótese que el algoritmo de selección de muestra utilizado debe contemplar la inclusión en todas las posibles muestras realizadas de todos los elementos del subgrupo de inclusión forzosa.

La selección de una muestra con diseño de muestreo Poisson se realiza mediante un algoritmo secuencial definido de manera similar que el algoritmo utilizado en la selección de muestras con diseño de muestreo Bernoulli.

1. Fijar para cada $k \in U$ el valor de la probabilidad de inclusión π_k tal que $0 < \pi_k \leq 1$.

2. Obtener ε_k para $k \in U$ como N realizaciones independientes de una variable aleatoria con distribución uniforme en el intervalo $[0, 1]$.

3. El elemento k-ésimo pertenece a la muestra con probabilidad π_k. Es decir, si $\varepsilon_k < \pi$ el individuo k-ésimo es seleccionado.

Dado que $\varepsilon_k \sim Unif[0, 1]$, se tiene que $Pr(\varepsilon_k < \pi_k) = \pi_k$ para $k \in U$. Por tanto, la inclusión de los individuos k-ésimo y l-ésimo, para $k \neq l$, es independiente; sin embargo, la distribución de $I_k(S)$ no es de tipo Binomial puesto que las variables aleatorias $I_k(S)$ no son idénticamente distribuidas.

Resultado 4.1.2. *Bajo muestreo Poisson, el tamaño de muestra $n(S)$ es una variable aleatoria, tal que*

$$E(n(S)) = \sum_U \pi_k \qquad Var(n(S)) = \sum_U \pi_k(1 - \pi_k) \tag{4.1.2}$$

Prueba. Utilizando el resultado 2.1.4 y las propiedades de una suma de cuadrados es suficiente probar que $\pi_{kl} = Pr(k \in S, l \in S) = \pi_k\pi_l$ para $k \neq l$, lo cual se tiene de inmediato dado que las variables aleatorias $I_k(S)$ e $I_l(S)$ son independientes. ■

Resultado 4.1.3. *Para el diseño de muestreo Poisson, las probabilidades de inclusión de primer y segundo orden están dadas por:*

$$\pi_k = \pi_k \tag{4.1.3}$$

$$\pi_{kl} = \begin{cases} \pi_k & \text{para } k = l \\ \pi_k\pi_l & \text{en otro caso} \end{cases} \tag{4.1.4}$$

respectivamente.

4.1.2 El estimador de Horvitz-Thompson

Resultado 4.1.4. *Para el diseño de muestreo Poisson, el estimador de Horvitz-Thompson, su varianza y su varianza estimada están dados por:*

$$\hat{t}_{y,\pi} = \sum_S \frac{y_k}{\pi_k} \tag{4.1.5}$$

$$Var_{PO}(\hat{t}_{y,\pi}) = \sum_U \left(\frac{1}{\pi_k} - 1\right) y_k^2 \tag{4.1.6}$$

$$\widehat{Var}_{PO}(\hat{t}_{y,\pi}) = \sum_S (1 - \pi_k) \left(\frac{y_k}{\pi_k}\right)^2 \tag{4.1.7}$$

respectivamente.

Prueba. Utilizando el resultado 2.2.2, se sigue que la demostración es inmediata puesto que

$$\Delta_{kl} = \begin{cases} \pi_{kl} - \pi_k\pi_l = \pi_k\pi_l - \pi_k\pi_l = 0 & \text{para } k \neq l \\ \pi_{kk} - \pi_k^2 = \pi_k(1 - \pi_k) & \text{para } k = l \end{cases} \tag{4.1.8}$$

luego la doble suma en la varianza del estimador de Horvitz-Thompson pasa a ser una sola suma. La demostración para el estimador de la varianza se lleva a cabo de manera análoga. ■

Ejemplo 4.1.2. Para nuestra población de ejemplo U, suponga que el individuo **Erik** debe estar en la muestra seleccionada; es decir, $\pi_{Erik} = 1$. Por tanto, existen $\binom{1}{1}2^4 = 16$ posibles muestras. Si el vector de probabilidades de inclusión para cada elemento de la población está dado por $(0.5, 0.2, 1, 0.9, 0.5)$. Realice el cálculo léxico-gráfico del estimador de Horvitz-Thompson y compruebe el insesgamiento, la varianza y las propiedades del diseño de muestreo.

4.1.3 Optimalidad en la estrategia de muestreo Poisson

Como se mencionó en capítulos anteriores, una estrategia de muestreo que utilice el estimador de Horvitz-Thompson, es óptima cuando las probabilidades de inclusión inducidas por el diseño de muestreo utilizado están correlacionadas positivamente con la característica de interés; en otras palabras, cuando $\pi_k \propto y_k$. En este caso utópico, y si se supone que el diseño de muestreo es de tamaño de muestra fijo ($n(S) = n$), el estimador de Horvitz-Thompson reproduciría el parámetro de interés t_y con varianza nula cuando las probabilidades de inclusión toman la siguiente forma $\pi_k = n\frac{y_k}{t_y}$. De esta forma, la estrategia utilizada sería una estrategia representativa con respecto a la variable de interés, puesto que para cualquier muestra seleccionada, el estimador de Horvitz-Thompson sería igual a t_y.

Resultado 4.1.5. *Suponiendo un tamaño de muestra fijo, bajo un diseño de muestreo Poisson, la varianza del estimador de Horvitz-Thompson se minimiza cuando*

$$\pi_k = \frac{ny_k}{\sum_U y_k} \tag{4.1.9}$$

Prueba. El objetivo es encontrar valores de π_k, tales que $0 < \pi_k \leq 1$ que minimicen la varianza del estimador de Horvitz-Thompson bajo diseño de muestreo Poisson, lo anterior se tiene cuando se realiza un censo, es decir cuando $\pi_k = 1$ para todo $k \in U$. Sin embargo, en la práctica se desea seleccionar una muestra de tamaño menor a N. Por tanto, minimizar $Var_{PO}(\hat{t}_{y,\pi})$ es equivalente a minimizar $\sum_U \frac{y_k^2}{\pi_k}$ sujeto a la restricción de un tamaño de muestra fijo, tal que $\sum_U \pi_k = n$. Luego la cantidad a minimizar está dada por el siguiente producto

$$\left(\sum_U \frac{y_k^2}{\pi_k}\right)\left(\sum_U \pi_k\right)$$

Una solución al anterior problema es utilizar la desigualdad de Cauchy-Schwartz, por tanto

$$\left(\sum_U \frac{y_k^2}{\pi_k}\right)\left(\sum_U \pi_k\right) \geq \left(\sum_U y_k\right)^2$$

Con igualdad cuando $\dfrac{y_k}{\pi_k} = c$, con c una constante. Ahora, se tiene que

$$n = \sum_U \pi_k = \sum_U \frac{y_k}{c}$$

Luego,

$$c = \sum_U \frac{y_k}{n}$$

Por tanto,

$$\pi_k = \frac{ny_k}{\sum_U y_k}$$



El anterior resultado es una ambigüedad puesto que con esa escogencia de las probabilidades de inclusión se asume que la característica de interés es conocida para toda la población. Si lo anterior sucede, no existiría la necesidad de estimar t_y. Sin embargo, Särndal, Swensson & Wretman (1992) aseguran que como el diseño de muestreo Poisson es de tamaño de muestra variable es ineficiente y utilizar el anterior razonamiento implicaría que el estimador de Horvitz-Thompson tome la siguiente forma

$$\hat{t}_{y,\pi} = \sum_S \frac{y_k}{\pi_k} = \frac{t_y}{n}\sum_S 1 = t_y\frac{n(S)}{n} \tag{4.1.10}$$

Por tanto, la variación del estimador calculado en cada muestra estaría dada por la variación del tamaño de muestra esperado $n(S)$. El anterior razonamiento nos lleva a pensar que el estimador de Horvitz-Thompson tendría un excelente desempeño bajo diseños de muestreo tales que $\pi_k \propto y_k$ y que induzcan muestras de tamaño fijo. Por otro lado, si el marco de muestreo tiene la virtud de adjuntar información auxiliar continua, por medio de una característica de interés x_k (en otras palabras, conocer el vector de características auxiliares $x_1, x_2, ..., x_N$ antes de realizar el muestreo) que esté muy bien correlacionada con la variable de interés, entonces la varianza de la estrategia de muestreo sería mínima cuando

$$\pi_k = n\frac{x_k}{\sum_U x_k} \tag{4.1.11}$$

Por otro lado, y siguiendo el mismo razonamiento que en el diseño de muestreo Bernoulli, como se tiene un marco de muestreo de elementos, entonces se conoce el tamaño poblacional N. De esta manera, un estimador para el total poblacional de la característica de interés con menor varianza es el llamado estimador alternativo dado por la expresión (2.2.18), que para el caso particular de muestreo Poisson toma la siguiente forma

$$\hat{t}_{y,alt} = \hat{t}_{y,\pi}\frac{N}{\hat{N}_\pi} \tag{4.1.12}$$

Para estimar la media poblacional, es posible utilizar este mismo razonamiento y junto con la expresión (2.2.15) resulta un estimador menos disperso

$$\tilde{y}_S = \frac{\hat{t}_{y,\pi}}{\hat{N}_\pi} \tag{4.1.13}$$

La forma estructural de los anteriores estimadores es una razón, cociente de dos cantidades aleatorias, y así se reduce parte de la variabilidad del estimador de Horvitz-Thompson que viene del hecho de que el tamaño muestral no es fijo para este diseño.

4.1.4 Marco y Lucy

Aunque esta estrategia de muestreo no fue utilizada en el sentido práctico y tiene una varianza alta dado que el tamaño de muestra es variable, es posible obtener buenos resultados que incentivar el uso de las estrategias de muestreo con probabilidad proporcional al tamaño. En primer lugar, se debe suponer que el marco de

muestreo contiene una característica auxiliar continua que será usada en la etapa de diseño y selección de la muestra.

Raj (1968) señala que en el caso concreto de una población agrícola, una característica auxiliar puede ser el área cultivada, para el caso de hogares, una característica auxiliar puede ser el número de personas que habitan en el hogar. Lehtonen & Pahkinen (2003) dan ejemplos claros acerca de las características auxiliares en encuestas de empresas y afirman que para este caso particular una característica auxiliar comúnmente usada es el número de empleados en la empresa; para el caso de encuestas a escuelas, una característica auxiliar es el número de alumnos. En encuestas a hospitales Bautista (1998) afirma que una característica auxiliar es el número de camas por hospital, no así el número de pacientes, pues esta última característica tiene una variación alta y está ligada a la temporada de realización de la encuesta.

Recuérdese que se quieren estimar tres totales de las características de interés Ingreso, Empleados e Impuestos del último periodo fiscal en las empresas del sector industrial. Para efectos prácticos, suponga que el marco de muestreo contiene todos los registros de cada una de las empresas del sector industrial de la característica Ingreso; de esta manera se podrá estimar el total poblacional para las características Empleados e Impuestos. Para efectos académicos, se estimará el total poblacional de la característica Ingreso, resaltando que hacerlo es una ambigüedad porque si se conocen todos los valores poblacionales de la característica de interés no hay necesidad de estimar lo que ya es conocido; sin embargo, como ejercicio académico es completamente admisible.

Con los supuestos anteriores, el marco de muestreo se carga en el ambiente de programación de R, nótese que el marco de muestreo ahora contiene cinco columnas, cuatro que se refieren a la identificación y/o ubicación geográfica y una columna que contiene los registros para la característica Ingreso.

```
> data(Lucy)
> dim(Lucy)
[1] 2396    5
```

Las probabilidades de inclusión deben ser creadas y están dadas por (4.1.9). Nótese que se debe fijar un tamaño esperado de muestra. Para que los resultados sean comparables, se utilizará un tamaño esperado de muestra de $n(S) = 400$. Una vez que las probabilidades de inclusión para todas las empresas del sector industrial han sido creadas, se debe verificar que cada una de ellas sea menor a la unidad; para esto, se utiliza la función `which` que R trae implementada en su ambiente básico y cuya salida es un conjunto de índices para los cuales la instrucción dentro del paréntesis es verdadera; cuando no existe ningún índice que cumpla `(pik>1)`, la función arroja la siguiente salida `integer(0)`. Sin embargo, si hubiese existido algún registro para el cual la instrucción `(pik>1)` sea cierta, se deben convertir las respectivas probabilidades de inclusión en la unidad.

```
> attach(Lucy)
> N <- dim(Lucy)[1]
> n <- 400
> pik <- n*Income/sum(Income)
```

```
> which(pik>1)
integer(0)
> sum(pik)
[1] 400
```

Nótese que la suma de las probabilidades de inclusión es igual al tamaño de muestra esperado.

```
> cor(pik,cbind(Income, Employees, Taxes))
     Income Employees    Taxes
[1,]      1  0.645536 0.916954
```

La correlación entre las probabilidades de inclusión inducidas mediante este diseño de muestreo Poisson es buena. Por supuesto, la correlación entre las π_k y la variable ingreso es uno pues las primeras son función lineal de Ingreso. Ahora, la cantidad de impuestos que las empresas del sector industrial declaran en un año fiscal, es proporcional al ingreso de las mismas; de hecho, si una empresa tiene ganancias nulas, entonces declarará impuestos nulos. Por otro lado, aunque una empresa tenga ganancias nulas, no necesariamente tendrá cero empleados; de hecho, en el sector industrial existen casos en donde una empresa con pocos empleados, tiene ingresos más altos que una empresa con muchos empleados; sin embargo, esta particularidad no se presenta de manera general, si esto fuera así, la correlación sería negativa y la característica de auxiliar Ingreso no debería ser utilizada en la estimación del total de la característica de interés Empleados. La figura 4.1 muestra el diagrama de dispersión de las tres variables de interés contra el vector de probabilidades de inclusión.

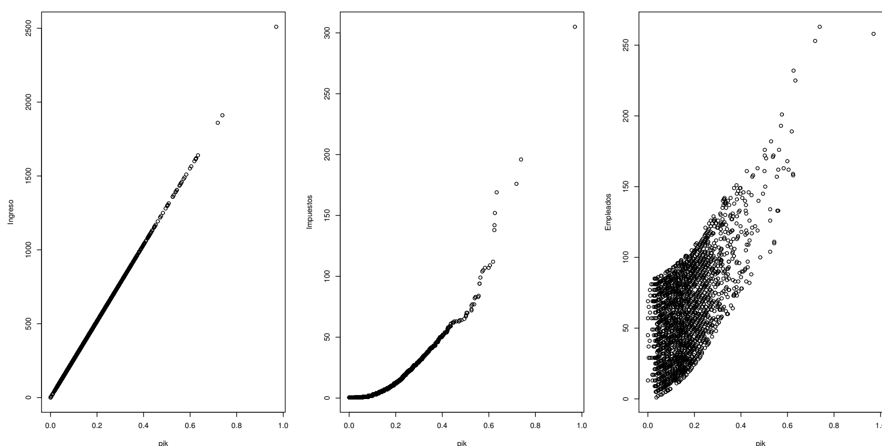


Figura 4.1: *Correlación de las probabilidades de inclusión con las características de interés.*

Para seleccionar la muestra bajo un diseño de muestreo Poisson, se utiliza la función `S.PO` del paquete `TeachingSampling`. Esta función consta de dos argumentos, `N`, el tamaño poblacional y `pik`, el vector de probabilidades de inclusión para cada elemento de la población. En nuestro caso, `pik` es el vector de probabilidades

creado anteriormente; pero, en general, puede ser utilizado cualquier vector de números entre cero y uno. La función `S.PO` devuelve un conjunto de índices que aplicados a la población resulta en los valores de las características de interés para cada miembro de la muestra seleccionada.

```
> sam <- S.PO(N,pik)
> muestra <- Lucy[sam,]
> attach(muestra)
> muestra

     Identificador Ubicación   Level Zona Income
3            AB003      c1k3 Pequeño    A    405
11           AB011     c1k11 Pequeño    A    374
13           AB013     c1k13 Pequeño    A    402
...
2394         AB986    c26k18  Grande    E   1297
2395         AB987    c26k19  Grande    E   1640

> n.s <- dim(muestra)[1]
> n.s
[1] 404
```

En este caso particular, la primera empresa seleccionada es la identificada con el número **AB003** y la última empresa en ser seleccionada es la identificada con el número **AB987**. Nótese que el marco de muestreo incluye la característica auxiliar Ingreso y que el tamaño efectivo de muestra es 404. Una vez que el trabajo de campo ha concluido, comienza la etapa de estimación, en donde se utilizará la función `E.PO` del paquete `TeachingSampling` que consta de dos argumentos, la matriz o vector de valores de la o las características de interés y `pik.s` los valores del vector de probabilidad de inclusión de cada uno de los elementos seleccionados en la muestra. En este caso particular se crea un conjunto de datos con la información muestral de las características de interés llamado `estima`. Nótese que la longitud del vector `pik.s` es de 404. La función `E.PO` devuelve las estimaciones del total poblacional, la varianza estimada y el respectivo coeficiente de variación de la(s) característica(s) de interés.

```
> pik.s <- pik[sam]
> estima <- data.frame(Income, Employees, Taxes)
> E.PO(estima,pik.s)
                                Income    Employees    Taxes
Total estimado            1.071450e+06 1.661678e+05 2.895542e+04
Varianza                  2.150761e+09 9.686064e+07 1.717237e+06
coeficiente de variación 4.328369e+00 5.922798e+00 4.525695e+00
```

La tabla 4.1. muestra los resultados particulares para esta estrategia de muestreo. Nótese que la característica Impuestos, tiene un menor coeficiente de variación porque está mucho mejor correlacionada con el vector de probabilidades de inclusión, mientras que la característica Empleados presenta un mayor coeficiente de variación. Desde un punto de vista completamente académico, está bien afirmar

que la estrategia de muestreo utilizada puede ser optimizada si se utiliza un diseño de muestreo con probabilidades de inclusión proporcionales al tamaño de alguna característica auxiliar, pero que induzca muestras de tamaño fijo. Nótese que, aunque el vector de probabilidades de inclusión tiene una correlación de uno con respecto a la característica Ingreso, el coeficiente de variación estimado para esta es de un 4.32 %, cifra que no es alta, pero que no paga el precio de utilizar esta información auxiliar en la etapa de diseño. Véase que los coeficientes de variación son un poco más bajos que al utilizar un diseño de muestreo Bernoulli, pero no más bajos que los obtenidos al usar un diseño de muestreo aleatorio simple.

Tabla 4.1: *Muestreo Poisson: estimación de los totales de las características de interés.*

Variable	Total poblacional	Total estimado	cve %	Desv. %
Ingreso	1035217	1071450	4.32	3.50
Empleados	151950	166167	5.92	9.36
Impuestos	28654	28955	4.52	1.05

La figura 4.2. muestra el comportamiento del estimador de Horvitz-Thompson para el total poblacional de la característica Ingreso. Se realizaron varios experimentos de Monte Carlo con el propósito de tener un examen más cercano del estimador de Horvitz-Thompson del total de la característica `Income` en la población `Lucy`. El resultado de la simulación se muestra en el histograma.

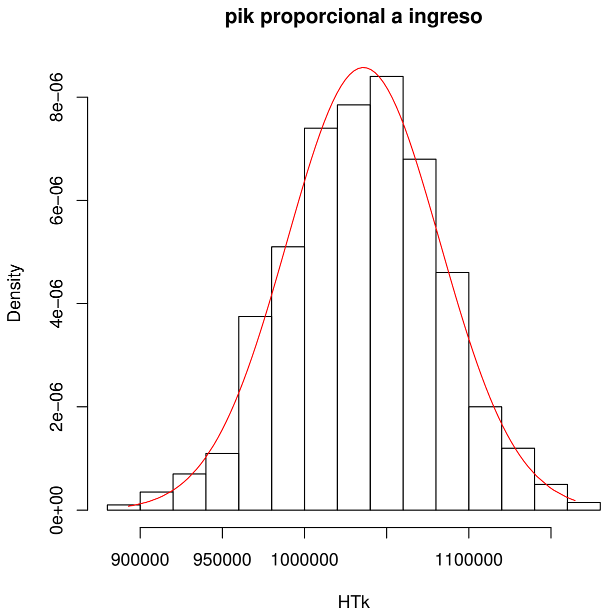


Figura 4.2: *Distribución muestral del estimador de Horvitz-Thompson bajo diseño Poisson.*

La media de las estimaciones de t_y es 1035851 que ajusta bien con el parámetro correspondiente $t_y = 1035217$. La distribución parece ser simétrica con forma de

campana (los valores de la distribución teórica se muestran en la curva sólida y roja) y no se notan grandes discrepancias entre lo observado y lo teórico. Sin embargo, al comparar la distribución de este estimador, se puede afirmar que la estrategia es más precisa que las anteriormente utilizadas pues se nota que el rango de valores que toma el estimador es menor que bajo otros diseños de muestreo.

4.2 Diseño de muestreo PPT

Siguiendo con el razonamiento que se introdujo en la sección anterior, Bautista (1998) afirma que en un diseño de muestreo con reemplazo, los valores óptimos de las probabilidades de selección para cada elemento de la población tendrían que estar dados por

$$p_k = \frac{y_k}{t_y}.$$

Con esta escogencia, el estimador de Hansen-Hurwitz estimaría al total poblacional de la característica de interés con varianza nula. De otra forma, el tamaño de muestra necesario para obtener una estimación con sesgo nulo sería de $m = 1$. Nótese que por (2.2.34), el estimador de Hansen-Hurwitz, es un promedio de m estimaciones. Con la escogencia de probabilidades de selección anterior, y con un tamaño de muestra de $m = 1$, se tiene que

$$\begin{aligned}
\hat{t}_{y,p} &= \frac{1}{1}\sum_{i=1}^{1}\frac{y_{k_i}}{p_{k_i}}\\
&= \frac{y_{k_i}}{p_{k_i}}\\
&= t_y\frac{y_{k_i}}{y_{k_i}} = t_y
\end{aligned}$$

Por supuesto, desde el punto de vista práctico sería una vez más, una ambigüedad la escogencia de las anteriores probabilidades de selección. Sin embargo, si el marco de muestreo es tal que contiene el valor de una característica continua auxiliar x_k bien relacionada con la característica de interés y_k para cada elemento de la población, es posible mediante el estimador de Hansen-Hurwitz, estimar el parámetro de interés con una varianza pequeña. De hecho, entre mejor correlación exista entre y_k y x_k menor varianza tendrá el estimador de Hansen-Hurwitz.

Definición 4.2.1. *Sea x_k, el valor de una característica auxiliar continua para el elemento k-ésimo tal que:*

1. *$x_k > 0$ para todo $k \in U$ y*
2. *x_k está disponible y es conocida para todos los elementos de la población.*

Entonces, se define un diseño de muestreo con probabilidad de selección proporcional al tamaño de la característica auxiliar, de la siguiente manera

$$p(s) = \begin{cases} \frac{m!}{n_1(s)!\ldots n_N(s)!} \prod_U \left(\frac{1}{p_k}\right)^{n_k(s)} & si \sum_U n_k(s) = m \\ 0 & en\ otro\ caso \end{cases} \tag{4.2.1}$$

Donde $n_k(s)$ es el número de veces que el elemento k-ésimo es seleccionado en la muestra realizada s y p_k es la probabilidad de selección del elemento k-ésimo dada por

$$p_k = \frac{x_k}{t_x}. \tag{4.2.2}$$

con t_x el total poblacional de la característica auxiliar x.

Resultado 4.2.1. *Para este diseño de muestreo, el soporte Q tiene cardinalidad igual a*

$$\#(Q) = \binom{N+m-1}{m}$$

Resultado 4.2.2. *Dado el soporte Q, de todas las posibles muestras con reemplazo de tamaño m, se verifica que el diseño de muestreo con probabilidad de selección proporcional al tamaño de la característica auxiliar es tal que*

$$\sum_{s\in Q} p(s) = 1$$

Prueba. Dado que

$$\sum_U p_k = \sum_U \frac{x_k}{t_x} = 1$$

entonces la demostración del resultado es inmediata haciendo uso del teorema multinomial. ■

Resultado 4.2.3. *Para un diseño de muestreo con reemplazo y con probabilidades de selección proporcionales al tamaño de una característica de información auxiliar, las probabilidades de inclusión de primer y segundo orden están dadas por*

$$\pi_k = 1-(1-p_k)^m \tag{4.2.3}$$
$$\pi_{kl} = 1-(1-p_k)^m-(1-p_l)^m+(1-p_k-p_l)^m \tag{4.2.4}$$

respectivamente. En donde $p_k = \frac{x_k}{t_x}$

Prueba. Utilizando el resultado 2.2.9 se llega a la demostración inmediata. ■

Cuando se tienen las cantidad del resultado 3.3.3, se pueden implementar los principios del estimador de Horvitz-Thompson para estimar el total poblacional t_y; sin embargo, el cálculo y estimación de la varianza de esta estrategia de muestreo resulta ser muy compleja computacionalmente.

4.2.1 Algoritmo de selección

Método acumulativo total

Hansen, Hurwitz & Madow (1953) plantearon este método de selección para ser utilizado junto con el estimador que lleva su nombre. Este método es conocido con el nombre de **algoritmo acumulativo total** y consiste en m selecciones independientes de tamaño 1, tal que:

- Sea
$$p_k = \frac{x_k}{t_x} \tag{4.2.5}$$
- Sea
$$T_k = \sum_{l=1}^{k} x_l \tag{4.2.6}$$
con $T_0 = 0$
- Obtener ε como una realización de una variable aleatoria con distribución uniforme en el intervalo (0,1).
- Seleccionar el k-ésimo elemento si $T_{k-1} < \varepsilon T_N \leq T_k$.

Al repetir m veces el anterior procedimiento, se ha seleccionado una muestra de un diseño con reemplazo con probabilidades de selección son proporcionales al tamaño de la característica de interés. Como este diseño de muestreo es con reemplazo, cuando existan elementos en la población cuyo valor de la característica auxiliar es muy grande, éstos elementos podrán ser seleccionados muchas veces porque sus probabilidades de selección son grandes con respecto a los demás elementos.

Método de Lahiri

En algunas ocasiones, cuando el tamaño poblacional N es muy grande, el anterior método resulta ineficiente. Lahiri (1951) plantea el siguiente algoritmo de selección: Siendo $M \geq \text{máx}(x_1, \ldots, x_N)$, los siguientes dos pasos se ejecutan para seleccionar un elemento.

1. Seleccione un número l de manera aleatoria de una distribución de probabilidad uniforme discreta en el intervalo $[1, N]$.
2. Seleccione un número η de manera aleatoria de una distribución de probabilidad uniforme discreta en el intervalo $[1, M]$.

Si $\eta \leq x_l$, entonces el elemento l-ésimo es seleccionado. Si, por el contrario, $\eta > x_l$ se repite el procedimiento hasta seleccionar una unidad. Si el tamaño de la muestra a seleccionar es m, entonces el anterior esquema se realiza m veces.

Ejemplo 4.2.1. Suponga que para la población de ejemplo U se tiene conocimiento de cada valor de la siguiente característica de información auxiliar correlacionada con la característica de interés.

```
> x <- c(52,60,75,100,50)
> x
[1] 52 60 75 100 50
```

Para seleccionar una muestra con probabilidad proporcional a `x`, se crean las probabilidades de selección dadas por

```
> pk <- x/sum(x)
> pk
[1] 0.1543027 0.1780415 0.2225519 0.2967359 0.1483680
```

Para seleccionar una muestra con reemplazo de la población U mediante el método acumulativo total, el paquete `TeachingSampling` implementa la función `S.PPS` que consta de dos argumentos, `m` el tamaño de muestra y `x` la característica de interés que contiene todos y cada uno de los valores correspondientes a los elementos de la población para la característica auxiliar.

```
> sam <- S.PPS(3,x)
> U[sam]
[1] "Sharon" "Sharon" "Leslie"
```

La salida de la función `S.PPS` es un conjunto de índices (no necesariamente distintos) que aplicados a los rótulos poblacionales proporcionan la muestra seleccionada.

4.2.2 El estimador de Hansen-Hurwitz

Hansen & Hurwitz (1943) propusieron el siguiente estimador insesgado para el parámetro de interés t_y con ayuda de información auxiliar continua en la etapa de diseño.

Resultado 4.2.4. *Sea x_k, el valor de una característica auxiliar continua, para un diseño de muestreo aleatorio proporcional al tamaño con reemplazo, el estimador de Hansen-Hurwitz del total poblacional t_y, su varianza y su varianza estimada están dados por:*

$$\hat{t}_{y,p} = \frac{t_x}{m}\sum_{i=1}^{m}\frac{y_{ki}}{x_{ki}} \tag{4.2.7}$$

$$Var_{PPT}(\hat{t}_{y,p}) = \frac{1}{m}\sum_{k=1}^{N}p_k\left(\frac{y_k}{p_k} - t_y\right)^2 \tag{4.2.8}$$

$$\widehat{Var}_{PPT}(\hat{t}_{y,p}) = \frac{1}{m(m-1)}\sum_{i=1}^{m}\left(\frac{y_i}{p_i} - \hat{t}_{y,p}\right)^2 \tag{4.2.9}$$

respectivamente, con p_k dados por (4.2.2). Nótese que $\hat{t}_{y,p}$ es insesgado para el total poblacional t_y de la característica de interés y, y que $\widehat{Var}_{MRAS}(\hat{t}_{y,p})$ es insesgado para $Var_{MRAS}(\hat{t}_{y,p})$.

Prueba.

$$
\begin{aligned}
E\left(\frac{t_x}{m}\sum_{i=1}^{m}\frac{y_{ki}}{x_{ki}}\right) &= E\left(\frac{t_x}{m}\sum_{U}n_k(S)\frac{y_k}{x_k}\right)\\
&= \frac{t_x}{m}\sum_{U}E(n_k(S))\frac{y_k}{x_k}\\
&= \frac{t_x}{m}\sum_{U}m\frac{x_k}{t_x}\frac{y_k}{x_k} = t_y
\end{aligned}
$$

dado que $E(n(S)) = mp_k$. Utilizando el resultado 2.2.13 y 2.2.14, se llega a la demostración de las varianzas. ■

Resultado 4.2.5. *Para el diseño de muestreo PPT, el estimador de Hansen-Hurwitz del total de la característica de información auxiliar reproduce ese total con varianza nula*

Prueba. De la definición del estimador Hansen-Hurwitz, y de la expresión (4.2.2), se tiene que

$$
\hat{t}_{x,p} = \frac{1}{m}\sum_{k\in S}\frac{x_k}{p_k} = \frac{1}{m}\sum_{k\in S}t_x = t_x
$$

Por otro lado,

$$
\begin{aligned}
Var_{PPT}(\hat{t}_{y,p}) &= \frac{1}{m}\sum_{k=1}^{N}p_k\left(\frac{x_k}{p_k}-t_x\right)^2 &(4.2.10)\\
&= \frac{1}{m}\sum_{k=1}^{N}p_k(t_x-t_x)^2 = 0 &(4.2.11)
\end{aligned}
$$

con lo cual se concluye la demostración ■

Resultado 4.2.6. *La varianza del estimador de Hansen-Hurwitz también puede ser escrita como*

$$
Var_{PPT}(\hat{t}_{y,p}) = \frac{1}{m}\sum_{U}\sum_{k<l}p_kp_l\left(\frac{y_k}{p_k}-\frac{y_l}{p_l}\right)^2 \qquad (4.2.12)
$$

Prueba. Desarrollando términos, se tiene que

$$
\begin{aligned}
\frac{1}{m}\sum_U\sum_{k<l}p_kp_l\left(\frac{y_k}{p_k}-\frac{y_l}{p_l}\right)^2 &= \frac{1}{2m}\sum\sum_{k,l}p_kp_l\left(\frac{y_k}{p_k}-\frac{y_l}{p_l}\right)^2\\
&= \frac{1}{2m}\sum_{k\in U}p_k\sum_{l\in U}p_l\left(\frac{y_k}{p_k}-\frac{y_l}{p_l}\right)^2\\
&= \frac{1}{2m}\sum_{k\in U}p_k\sum_{l\in U}\left(p_l\frac{y_k^2}{p_k^2}-2\frac{y_ky_l}{p_k}+\frac{y_l^2}{p_l}\right)\\
&= \frac{1}{2m}\sum_{k\in U}p_k\left(\frac{y_k^2}{p_k^2}-2\frac{y_k}{p_k}t_y+\sum_{l\in U}\frac{y_l^2}{p_l}\right)\\
&= \frac{1}{2m}\left(\sum_{k\in U}\frac{y_k^2}{p_k}-2t_y^2+\sum_{l\in U}\frac{y_l^2}{p_l}\right)\\
&= \frac{1}{m}\left(\sum_{k\in U}\frac{y_k^2}{p_k}-t_y^2\right)\\
&= \frac{1}{m}\sum_{k\in U}\left(\frac{y_k^2}{p_k}-p_kt_y^2\right)\\
&= \frac{1}{m}\sum_{k\in U}p_k\left(\frac{y_k^2}{p_k^2}-2\frac{y_k}{p_k}t_y+t_y^2\right)\\
&= \frac{1}{m}\sum_{k\in U}p_k\left(\frac{y_k}{p_k}-t_y\right)^2
\end{aligned}
$$

y esta última expresión coincide con la varianza del estimador de Hansen-Hurwitz en muestreo PPT. ■

Särndal, Swensson & Wretman (1992) afirman que la primera forma que toma la varianza y su estimación insesgada para el estimador de Hansen-Hurwitz es fácil de calcular computacionalmente. Sin embargo, la expresión alternativa de la varianza se utilizará para desarrollos teóricos posteriores.

Esta estrategia de muestreo es con reemplazo, y comparada con una estrategia de muestreo que utilice información auxiliar en la etapa de diseño con el estimador de Horvitz-Thompson es un poco menos eficiente. Sin embargo, en la práctica es más utilizada porque los cálculos computacionales son fáciles de realizar y es preferida porque con un número grande de elementos incluidos en la muestra, el cálculo de la varianza estimada del estimador de Horvitz-Thompson se hace inapropiado por la gran cantidad de productos cruzados.

La mayor pertinencia de esta estrategia está principalmente en la estimación de totales, como se verá más adelante surgen complicaciones, con respecto a la información auxiliar al usar un diseño de muestreo con reemplazo proporcional al tamaño en la estimación de razones. En encuestas de hogares, no resulta adecuado utilizar este diseño de muestreo, puesto que en una población, existe un número de hogares homogéneos por vivienda. Por otro lado, en encuestas de negocios y empresas es útil utilizar diseños proporcionales porque sí existen diferencias marcadas en los tamaños de las mismas; por ejemplo, en el número de empleados, el número

de metros cuadrados en las instalaciones, el ingreso, etc. La función de varianza para esta estrategia de muestreo no es monótona decreciente; por la configuración de la información auxiliar, la varianza puede aumentar cuando aumenta el tamaño de muestra.

Ejemplo 4.2.2. Para nuestra población de ejemplo U, existen $\binom{N+m-1}{m} = 20$ posibles muestras con reemplazo de tamaño $m = 2$. Utilizando la característica auxiliar x, realice el cálculo léxico-gráfico del estimador de Hansen-Hurwitz, compruebe el insesgamiento, calcule la varianza y el insesgamiento del estimador de la varianza.

4.2.3 Eficiencia de la estrategia

La regla de oro de una buena muestra reza que para que la inferencia basada en el diseño de muestreo arroje estimaciones que sean (abusando del lenguaje) de varianza mínima e insesgadas, las probabilidades de inclusión (o selección, según sea el caso) que arroje el diseño de muestreo utilizado deben ser directamente proporcionales a los valores que toma la característica de interés en la población. Raj (1954) demuestra el siguiente resultado que conduce condiciona el comportamiento estructural de la información auxiliar que debe cumplir dos condiciones para que la eficiencia de la estrategia PPT sea mayor que la del diseño aleatorio simple con reemplazo.

Resultado 4.2.7. *La resta de la varianza de la estrategia aleatoria simple con reemplazo con la varianza de la estrategia PPT da como resultado la siguiente expresión:*

$$Var_{MRAS}(\hat{t}_{y,p}) - Var_{PPT}(\hat{t}_{y,p}) = \frac{N^2}{m} Cov\left(x, \frac{y^2}{x}\right) \tag{4.2.13}$$

Prueba. Utilizando la expresión general de la varianza (2.2.36) bajo cualquier diseño de muestreo con reemplazo se tiene que

$$\begin{aligned}
Var_{MRAS}\left(\hat{t}_{y,p}\right) - Var_{PPT}(\hat{t}_{y,p}) &= \frac{1}{m}\left[N\sum_{k=1}^{N} y_k^2 - t_y^2 - t_x \sum_{k=1}^{N} \frac{y_k^2}{x_k} + t_y^2\right] \\
&= \frac{1}{m}\left[\sum_{k=1}^{N} \frac{y_k^2}{x_k}(Nx_k - t_x)\right] \\
&= \frac{N}{m}\left[\sum_{k=1}^{N} \frac{y_k^2}{x_k}(x_k - \bar{x})\right] \\
&= \frac{N^2}{m} Cov\left(x, \frac{y^2}{x}\right)
\end{aligned}$$

La última igualdad se tiene puesto que

$$
\begin{aligned}
NCov\,(x,w) &= \sum_{k=1}^{N}(x_k-\bar{x})(w_k-\bar{w})\\
&= \sum_{k=1}^{N}(x_k-\bar{x})w_k-\bar{w}\sum_{k=1}^{N}(x_k-\bar{x})=\sum_{k=1}^{N}(x_k-\bar{x})w_k
\end{aligned}
$$

■

El anterior resultado indica que para que la estrategia de muestreo PPT sea más eficiente en términos de varianza que la estrategia de muestreo MRAS, además de que $p_k \propto x_k$, es necesario que la correlación entre $\left(x, \dfrac{y^2}{x}\right)$ sea positiva. Nótese que si la razón entre y y x es contante e igual a C, se tiene que

$$
\begin{aligned}
Cor\left(x,\frac{y^2}{x}\right) &= Cor\left(x,y\frac{y}{x}\right)\\
&= Cor\,(x,yC)\\
&= Cor\,(x,y)
\end{aligned}
$$

Por tanto, una condición necesaria para que el diseño de muestreo PPT sea más eficiente que el diseño de muestreo MRAS es que exista una correlación positiva entre la característica de interés y la información auxiliar; pero, una condición suficiente para la optimalidad del diseño PPT, es que la razón $\frac{y_k}{x_k}$ permanezca constante para todo $k \in U$.

Además de la razón constante, Lehtonen & Pahkinen (2003) muestran que la eficiencia del diseño de muestreo PPT está directamente relacionada con el siguiente modelo de regresión

$$
y_k = \beta_0 + \beta_1 x_k + E_k \tag{4.2.14}
$$

que relaciona la característica de interés con la información auxiliar. Concluye que para que el diseño de muestreo PPT sea más eficiente que el diseño de muestreo MRAS, la cantidad β_0 debe ser pequeña. Es decir, que la línea de regresión ajuste cerca del origen. Es más, incluso si la correlación entre la característica de interés y la información auxiliar fuera perfecta e igual a uno, entonces no habría ningún término de error, pero aun así si β_0 es grande, entonces la estrategia de muestreo PPT podría arrojar una eficiencia menor a la del diseño de muestreo aleatorio simple con reemplazo.

La eficiencia de la estrategia de muestreo, depende de dos aspectos. Primero, el tipo de parámetro que se quiere estimar. Lehtonen & Pahkinen (2003) afirman que para la estimación de totales, la estrategia de muestreo PPT, funciona mejor, en términos de eficiencia, que para la estimación de razones o medianas. Segundo, que la razón entre x_k y y_k sea constante para toda la población.

4.2.4 Marco y Lucy

Una de las características del diseño de muestreo PPT es el uso de información auxiliar en la etapa de diseño. Obviamente, la información auxiliar debe estar presente en el marco de muestreo. En esta sección, de Marco y Lucy, seguiremos la tendencia que comenzamos en el diseño de muestreo Poisson. Suponga que, para todas las empresas del sector industrial, el valor del ingreso en el último año fiscal está disponible en el marco de muestreo.

Se quiere estimar, el total poblacional de las características de interés Empleados e Impuestos, para lo cual, se utilizará una estrategia de muestreo que utiliza un diseño de muestreo con reemplazo y probabilidades de selección de las empresas proporcionales al tamaño de la característica auxiliar Ingreso junto con el estimador de Hansen-Hurwitz. Como se vio antes, para que esta estrategia de muestreo sea óptima con respecto a una que utilice un diseño aleatorio simple con reemplazo se deben cumplir ciertas condiciones. Antes de analizarlas, veamos que, para este caso particular y con un tamaño de muestra igual a $m = 400$, el diseño de muestreo PPT es menos eficiente que el muestreo simple con reemplazo para la estimación del total de empleados, aunque es más eficiente que el muestreo simple con reemplazo para la estimación del total de impuestos declarados. Lo anterior se tiene utilizando la expresión (4.2.13) escrita en código de R.

```
> data(Lucy)
> attach(Lucy)
> N=2396
> m=400
> (N^2/m)*cov(Income, (Employees^2/Income))
[1] -37167215
> (N^2/m)*cov(Income, (Taxes^2/Income))
[1] 3586325
```

Primero, que la correlación entre `Income` y `y2/Income` sea positiva. Aunque la correlación entre `Income` y `Employees` e, `Income` y `Taxes` sea positiva, se debe verificar que la correlación entre `Income` y la nueva variable `Employees2/Income` sea positiva, como también la correlación entre `Income` y `Taxes2/Income`. Mediante el uso de la función `cor` que R incorpora en su ambiente de trabajo, se tiene que para la característica de interés Empleados, la correlación es negativa, aunque casi nula. Mientras que para la característica de interés Impuestos, la correlación buscada es positiva. Esto indica que para la estimación del total de empleados, el uso de la información auxiliar no conlleva a ganancias significativas en la eficiencia de la estrategia. Por otro lado, para la estimación del total de impuestos declarados, sí se tiene un ganancia significativa.

```
> cor(Income, (Employees^2/Income))
[1] -0.07727604
> cor(Income, (Taxes^2/Income))
[1] 0.7092853
```

Otra de las condiciones para la optimalidad de la estrategia es que el cociente entre `Income` y las características de interés `Taxes` y `Employees` sea constante para todo

elemento de la población. Mediante el uso de la función `plot` es posible tener un acercamiento gráfico al comportamiento de los respectivos cocientes. Nótese que la función `abline` permite trazar una línea sobre el promedio de los cocientes.

```
> par(mfrow=c(1,2))
> plot(Employees/Income)
> abline(h=mean(Employees/Income), col=2)
> plot(Taxes/Income)
> abline(h=mean(Taxes/Income), col=2)
```

La figura 4.3. muestra que la relación existente entre el cociente `Income` y `Employees` es uniforme en casi toda la población. Por supuesto, se observan algunos datos atípicos que están muy lejos de la línea de referencia, pero en general se observa un comportamiento homogéneo. Esto no ocurre con la relación existente entre el cociente `Income` e `Taxes` donde existe un comportamiento más disperso para todos los elementos de la población. A pesar de lo anterior, se puede afirmar que el comportamiento de la razón es constante.

Un tercer argumento para el uso de la estrategia de muestreo PPT es el examen del ajuste de una línea de regresión entre `Employees` con `Income` y `Taxes` con `Income` respectivamente. Para esto, se ajustan dos modelos. El primero dado por

$$Impuestos_k = \beta_0 + \beta_1 Ingreso + E_k \tag{4.2.15}$$

Para la estimación del total de la característica Impuestos y, el segundo dado por

$$Empleados_k = \beta_0 + \beta_1 Ingreso + E_k \tag{4.2.16}$$

Para la estimación del total de la característica Empleados. Para los modelos anteriores, nos interesa conocer el valor que toma el intercepto de cada línea de regresión. Si el intercepto β_0 es cercano a cero, entonces se ha ganado eficiencia al utilizar un diseño de muestreo PPT. R incorpora la función `lm` para el ajuste de modelos lineales. Las estimaciones de β_0 y β_1 se hacen por medio del método de los mínimos cuadrados. Un análisis de regresión de `y` contra `x` es especificado mediante `y ~x`. La salida de la función `lm` está dada por las estimaciones de los coeficientes de los modelos de regresión. Con ayuda de la función `summary` es posible extraer más información respecto a la inferencia de las estimaciones.

Para el primer modelo, se nota que la estimación del intercepto está dada por -13.77 y, a juzgar por las tres estrellas, es una cantidad significativa. Aunque para nuestro análisis está cerca del origen, por tanto se gana en eficiencia al utilizar esta estrategia de estimación para el total poblacional de la característica de interés Impuestos.

```
> M.I <- lm(Taxes~Income)
> summary(M.I)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.377e+01  2.689e-01  -51.19   <2e-16 ***
Income       5.954e-02  5.295e-04  112.45   <2e-16 ***
```

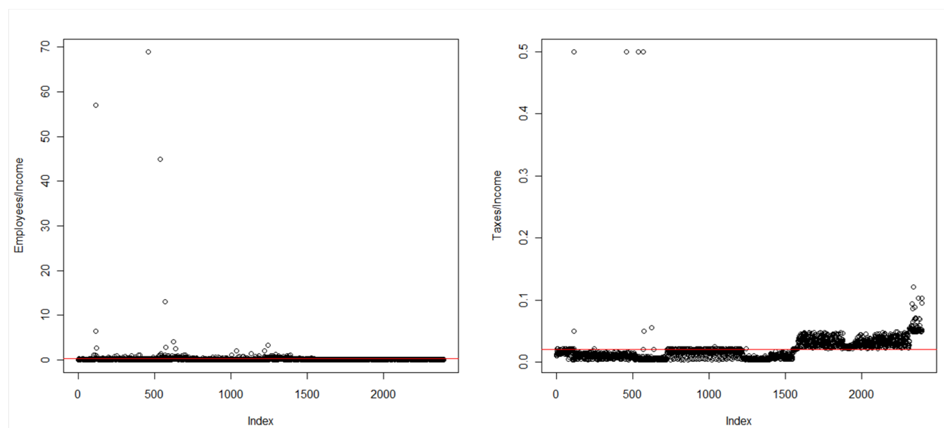


Figura 4.3: *Comportamiento del cociente de la información auxiliar con las características de interés.*

El intercepto del segundo modelo ha sido estimado como 29.06, a diferencia del modelo anterior, no se puede decir que está cerca del origen. Además, por la magnitud de la escala de medición de las características, se puede decir que es una cantidad importante y no despreciable.

```
> M.E <- lm(Employees~Income)
> summary(M.E)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 29.058392   0.976584   29.75   <2e-16 ***
Income       0.079525   0.001923   41.36   <2e-16 ***
```

La figura 4.4. muestra la línea de regresión ajustada para los dos modelos anteriores; es claro que el intercepto del modelo con impuestos declarados se puede considerar nulo, pero el intercepto del modelo con número de empleados es grande. Los tres anteriores argumentos permiten estar confiados al utilizar la estrategia de muestreo PPT para la estimación del total de impuestos declarados, pero se sabe que para la estimación del total de número de empleados, este diseño muestral no es más eficiente que el diseño simple con reemplazo.

Una vez se ha decidido usar la estrategia de muestreo PPT, es necesario seleccionar la muestra. En este caso, se ha querido utilizar el mismo tamaño de muestra, que en las anteriores estrategias de muestreo. En primer lugar, se adjunta el marco de muestreo que no sólo contiene la ubicación e identificación sino además el valor de la información auxiliar Ingreso para cada empresa del sector industrial. La selección de la muestra se hace mediante el uso de la función S.PPS para la cual los argumentos introducido son `m=400` junto con la información auxiliar `Income`. Esta función utiliza el algoritmo de selección acumulativo total.

```
> pk <- Income/sum(Income)
> sam <- S.PPS(m,Income)
```

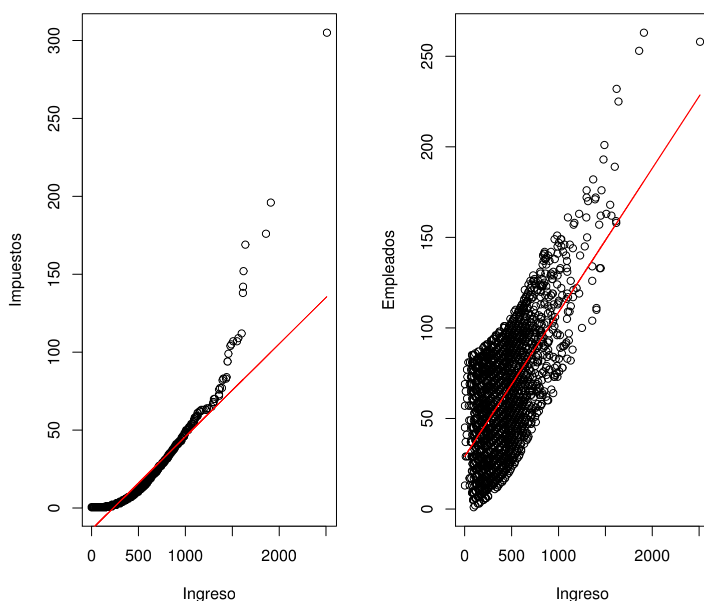


Figura 4.4: *Líneas de regresión.*

```
> muestra <- Lucy[sam,]
> attach(muestra)
> muestra
            ID Ubication  Level Zone Income
2148     AB804    c23k69 Medium    D    870
1579     AB058    c16k94 Medium    A    556
1780    AB1218    c18k97 Medium    B    580
...
2152     AB808    c23k73 Medium    D    986
2008.1   AB632    c22k28 Medium    D    710
```

El método acumulativo total no tiene en cuenta ningún ordenamiento. En este caso particular, la última empresa en ser seleccionada fue la empresa con número de identificación **AB632**, aunque esta empresa ya había sido seleccionada en la muestra en dos ocasiones. Es decir, fue seleccionada en tres ocasiones.

Una vez seleccionada la muestra con reemplazo, se utiliza la función `E.PPS` del paquete `TeachingSampling` cuyos argumentos son la(s) característica(s) de interés y un vector de probabilidades de selección `pk`. Por supuesto, el vector de probabilidades de selección en la población está dado por `pk <- Income/sum(Income)`. Sin embargo, en la función `E.PPS`, el vector de probabilidades debe corresponder a las probabilidades de selección de cada uno de los elementos elegidos en la muestra. En este caso la longitud del vector `pk.s` es de `m=400`.

```
> pk.s <- pk[sam]
```

```
> estima <- data.frame(Income, Employees, Taxes)

> E.PPS(estima,pk.s)
                                 Income    Employees        Taxes
Total estimado             1.035217e+06 1.489366e+05 2.860961e+04
Varianza                   4.643047e-27 4.859960e+03 4.970087e+02
coeficiente de variación 6.582186e-18 4.680745e-02 7.792379e-02
```

Los resultados de aplicar la estrategia de muestreo son muy favorables. Nótese, que a diferencia de la estrategia de muestreo Poisson, el total poblacional de la característica auxiliar ingreso, es estimada exactamente con varianza casi nula. El total poblacional de las características de interés Empleados e Impuestos tienen coeficientes de variación menores a 1 %. La tabla 4.2 muestra los resultados obtenidos en este ejercicio particular.

Tabla 4.2: *Muestreo PPT: estimación de los totales de las características de interés.*

Variable	Total poblacional	Total estimado	cve %	Desv. %
Ingreso	1035217	1035217	0.000	0.000
Empleados	151950	148936	0.004	-1.98
Impuestos	28654	28609	0.008	-0.16

Véase que también es posible estimar el tamaño de la población. Es decir, el número total de empresas que pertenecen al sector industrial. Con la configuración de las probabilidades de inclusión, lo único que se debe hacer es asignar un vector de unos, de tamaño `m=400` para obtener la estimación. En este caso, está dada por $\hat{N} = 2272$. Siendo que en la realidad el tamaño de la población en el sector industrial es de $N = 2396$.

```
> E.PPS(rep(1,times=400),pk.s)
                                      y
Total estimado             2.271888e+03
Varianza                   1.810680e+00
coeficiente de variación 5.922894e-02
```

Asimismo, una estrategia alternativa es utilizar un diseño de muestreo con reemplazo y probabilidad de selección proporcional al tamaño junto con el estimador de Horvitz-Thompson, el cual es también insesgado. Särndal, Swensson & Wretman (1992) se preguntan cuál es el mejor estimador y llegan a la conclusión que dependiendo de la configuración de los valores de las características de interés y de información auxiliar un estimador tendrá menor varianza que el otro. Por tanto, no es posible generalizar. De lo que sí se puede estar seguro, es de la simplicidad, en materia de cálculos del estimador de Horvitz-Thompson. En la práctica, este es un argumento muy fuerte que incentiva el uso del estimador de Hansen-Hurwitz.

Utilizando el resultado 4.2.3., es posible estimar los parámetros de interés mediante el uso del estimador de Horvitz-Thompson. Para esto, se calculan las probabilidades inclusión. Nótese que la suma de éstas es de 358. Se extraen las probabilidades de inclusión de los elementos en la muestra y se utiliza la forma genérica del estimador de Horvitz-Thompson.

```
> pik <- 1-(1-pk)^400
> sum(pik)
[1] 358.1504
> pik.s <- pik[sam]
> sum(1/pik.s)
[1] 2708.389
> colSums(estima/pik.s)
   Income  Employees  Taxes
1157132.18  166105.50   32318.71
```

Las estimaciones resultantes no son mejores, en el sentido práctico, a las obtenidas mediante el uso del estimador de Hansen-Hurwitz. Ahora, la estimación de la varianza supondría un esfuerzo computacional demasiado grande.

4.3 Diseño de muestreo πPT

Como se vio en la sección anterior, utilizar un esquema de muestreo con probabilidades proporcionales a alguna característica de información auxiliar puede resultar en ganancia de precisión. Sin embargo, utilizar una estrategia de muestreo que contemple un diseño de muestreo con reemplazo es menos eficiente que implementar una estrategia de muestreo que contemple un diseño de muestreo sin reemplazo y de tamaño muestral fijo.

En la sección anterior, se utilizó un diseño de muestreo con probabilidades proporcionales, con reemplazo y, sin embargo, arrojó muy buenos resultados en términos de eficiencia comparado con los diseños de muestreo de probabilidades simples. Esta sección se concentra en la implementación de diseños de muestreo con probabilidades de inclusión proporcionales a una característica de interés y cuya estructura general sea sin reemplazo. De esta forma, es posible aumentar dramáticamente la eficiencia de la estrategia que involucra al estimador de Horvitz-Thompson.

Lohr (2000) afirma que el muestreo de probabilidades simples, proporciona esquemas que, frecuentemente, son fáciles de explicar y diseñar. Sin embargo, estos esquemas no siempre pueden ser realizados puesto que las probabilidades simples no siempre reflejan el comportamiento de la característica de interés en la población.

Este diseño de muestreo induce probabilidades de inclusión proporcionales al tamaño de una característica de información auxiliar[1]. De esta manera, se supone que el marco de muestreo tiene la bondad de poseer información auxiliar de tipo continuo y positiva disponible para todo elemento perteneciente a la población finita. Asimismo, el diseño de muestreo πPT[2], de tamaño de muestra fijo e igual a N, se basa en la construcción de probabilidades de inclusión que obedezcan la siguiente relación:

$$\pi_k = \frac{nx_k}{t_x} \qquad 0 < \pi_k \leq 1 \tag{4.3.1}$$

[1] El requisito indispensable de la información auxiliar es que sea aproximadamente proporcional a la característica de interés.

[2] Nótese que la sigla πPT se refiere a los diseños de muestreo que inducen probabilidades de inclusión proporcionales a una característica de información auxiliar.

Además se busca que:

- El algoritmo de selección de muestras bajo este diseño sea de fácil implementación computacional.
- Las probabilidades de inclusión de segundo orden sean positivas, $\pi_{kl} > 0$. De lo contrario el estimador de la varianza podría ser sesgado.
- El cálculo de estas probabilidades de inclusión de segundo orden, π_{kl}, sea sencillo.
- $\Delta_{kl} < 0 \quad \forall k \neq l$ para que la estimación de la varianza no sea negativa.

Este diseño de muestreo se puede considerar como una generalización de la mayoría de diseños de muestreo sin reemplazo. Por ejemplo: si la característica de información auxiliar es constante e igual a C, entonces para un tamaño de muestra fijo, las probabilidades de inclusión de primer orden estarían dadas por:

$$\begin{aligned}\pi_k &= \frac{nx_k}{t_x}\\ &= \frac{nC}{NC} = \frac{n}{N}\end{aligned}$$

Con lo que se tiene un diseño de muestreo caracterizado por probabilidades simples. En ciertas ocasiones, cuando las población tiene un comportamiento muy variable, irregular y sesgado, algunas de las pi_k inducidas por la expresión (4.3.1) pueden ser mayores a uno para ciertos elementos. En tal caso, estos elementos son incluidos en todas las posibles muestras y toman el nombre de **elementos de inclusión forzosa**. Sin embargo, para calcular la probabilidad de inclusión de los elementos restantes, se debe excluir estos elementos de inclusión forzosa y volver a calcular las probabilidades de inclusión mediante una reformulación de la expresión (4.3.1) dada por

$$\pi_k = \frac{(n - n^*)x_k}{\sum_{k\in U^*} x_k} \qquad 0 < \pi_k \leq 1; \quad k \in U^* \tag{4.3.2}$$

donde n^* corresponde al número de elementos de inclusión forzosa y U^* la población finita excluyendo a estos elementos de inclusión forzosa. Al final del proceso, deberían existir dos grupos de elementos:

1. Un grupo de elementos de inclusión forzosa con probabilidades de inclusión iguales a uno.
2. Un grupo de elementos con probabilidades de inclusión $0 < \pi_k < 1$ y proporcionales a x_k.

Por tanto, el problema se reduce a la selección de n unidades con probabilidades de inclusión tales que

$$\sum_{k\in U} \pi_k = n$$

El siguiente resultado da cuenta de la forma estructural que toma el estimador de Horvitz-Thompson, de su varianza y de su varianza estimada.

Resultado 4.3.1. *Para el diseño de muestreo πPT, el estimador de Horvitz-Thompson, su varianza y su varianza estimada están dados por:*

$$\hat{t}_{y,\pi} = \sum_S \frac{y_k}{\pi_k} \tag{4.3.3}$$

$$Var_{\pi PT}(\hat{t}_{y,\pi}) = -\frac{1}{2}\sum\sum_U \Delta_{kl}\left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2 \tag{4.3.4}$$

$$\widehat{Var}_{\pi PT}(\hat{t}_{y,\pi}) = -\frac{1}{2}\sum\sum_S \frac{\Delta_{kl}}{\pi_{kl}}\left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2 \tag{4.3.5}$$

Resultado 4.3.2. *Para el diseño de muestreo πPT, el estimador de Horvitz-Thompson del total de la característica de información auxiliar reproduce ese total con varianza nula*

Prueba. De la definición del estimador de Horvitz-Thompson, y de la expresión (4.3.1),se tiene que

$$\hat{t}_{x,\pi} = \sum_{k\in S} \frac{x_k}{\pi_k} = \sum_{k\in S} t_x \frac{1}{n} = t_x$$

Por otro lado,

$$Var_{\pi PT}(\hat{t}_{x,\pi}) = -\frac{1}{2}\sum\sum_U \Delta_{kl}\left(\frac{x_k}{\pi_k} - \frac{x_l}{\pi_l}\right)^2 \tag{4.3.6}$$

$$= -\frac{1}{2}\sum\sum_U \Delta_{kl}\left(\frac{t_x}{n} - \frac{t_x}{n}\right)^2 = 0 \tag{4.3.7}$$

con lo cual se concluye la demostración ■

Ejemplo 4.3.1. Suponga que para la población de ejemplo U se tiene conocimiento de cada valor de la siguiente característica de información auxiliar correlacionada con la característica de interés. Por tanto, un primer paso para el cálculo de las probabilidades de inclusión es aplicar la expresión (4.3.1).

```
> n <- 4
> x <- c(52,60,75,100,50)
> pik <- n*x/sum(x)
> pik
[1] 0.6172107 0.7121662 0.8902077 1.1869436 0.5934718
```

Nótese que el cuarto elemento de la población, correspondiente a **Sharon** es un elemento de inclusión forzosa; es decir que está presente en todas las posibles muestras. El siguiente paso es separar a **Sharon** de los restantes elementos y proseguir con el cálculo de las probabilidades de inclusión inducidas por la expresión (4.3.2)

```
> n <- 3
> x <- c(52,60,75,50)
> pik <- n*x/sum(x)
> pik
[1] 0.6582278 0.7594937 0.9493671 0.6329114
```

Por tanto el vector de probabilidades de inclusión para toda la población U está dado por

$$\boldsymbol{\pi} = (\underbrace{0.6582278}_{\textbf{Yves}}, \underbrace{0.7594937}_{\textbf{Ken}}, \underbrace{0.9493671}_{\textbf{Erik}}, \underbrace{1.0000}_{\textbf{Sharon}}, \underbrace{0.6329114}_{\textbf{Leslie}})'$$

4.4 Selección de muestras πPT

Existen varios métodos de selección de muestras πPT. Sin embargo, todos ellos están basados en una teoría fuerte y complicada y, en algunas ocasiones, son muy difíciles de implementar en la práctica. A continuación, se exponen dos métodos de selección de muestras de tamaño $n = 1$ y $n = 2$. Särndal, Swensson & Wretman (1992) comentan que a simple vista parecería irreal considerar tamaños de muestra tan pequeños. Sin embargo, en muestreo estratificado y muestreo para conglomerados (ver siguientes capítulos) tiene sentido seleccionar solamente una o dos unidades primarias de muestreo.

Tamaño de muestra $n = 1$

Para $n = 1$ se utiliza el método acumulativo total, que consiste en:

1. Definir $T_0 = 0$ y $T_k = T_{k-1} + x_k$ $(k \in U)$.
2. Calcular un número aleatorio ε con distribución uniforme en el intervalo $[0, 1]$.
3. Si $T_{k-1} < \varepsilon T_N < T_k$, el elemento k-ésimo se selecciona.

Nótese que este algoritmo de selección garantiza que el diseño de muestreo es un autentico πPT puesto que

$$\pi_k = Pr(k \in S) = Pr(T_{k-1} < \varepsilon T_N < T_k) = \frac{T_k - T_{k-1}}{T_N} = \frac{x_k}{t_x}$$

Por supuesto, no es posible obtener un estimador insesgado de la varianza del estimador de Horvitz-Thompson puesto que la muestra sólo considera la inclusión de un elemento de la población finita.

Tamaño de muestra $n = 2$

En este escenario es preciso garantizar que las probabilidades de inclusión de primer orden estén dadas por

$$\pi_k = \frac{2x_k}{t_x}$$

para todo elemento de la población finita. En este caso, los dos elementos de la muestra son seleccionados uno por uno. Para tal fin, se debe seguir el siguiente algoritmo (Brewer 1963, Brewer 1975) que utiliza el método acumulativo total en cada una de las dos selecciones, así:

1. En la primera extracción, el elemento k-ésimo es seleccionado con probabilidad

   $$p_k = \frac{c_k}{\sum_{k \in U} c_k}$$

   donde

   $$c_k = \frac{x_k(T_N - x_k)}{T_N(T_N - 2x_k)}$$

2. En la segunda extracción, el elemento seleccionado en el paso anterior, digamos el elemento k^*, es retirado del sorteo. El segundo elemento es seleccionado con probabilidad

   $$p_{l|k^*} = \frac{x_l}{T_N - x_{k^*}}$$

Resultado 4.4.1. *Bajo el esquema de selección de Brewer las probabilidades de inclusión de primer orden satisfacen la siguiente relación*

$$\pi_k = \frac{2x_k}{t_x}$$

Las probabilidades de inclusión de segundo orden están dadas por

$$\pi_{kl} = \frac{2x_k x_l}{T_N(\sum_{k \in U} c_k)} \frac{T_N - x_k - x_l}{(T_N - 2x_k)(T_N - 2x_l)}$$

Prueba. La probabilidad de inclusión de primer orden del k-ésimo elemento está

dada por

$$
\begin{aligned}
\pi_k &= Pr(k \in S)\\
&= Pr(k \text{ sea seleccionado en la primera extracción})\\
&+ Pr(k \text{ sea seleccionado en la segunda extracción})\\
&= p_k + p_{k|j}\sum_{\substack{j\in U\\ j\neq k}} p_j\\
&= \frac{x_k(T_N-x_k)/T_N(T_N-2x_k)}{D}\\
&+ \sum_{\substack{j\in U\\ j\neq k}} \frac{x_j(T_N-x_j)/T_N(T_N-2x_j)}{D}\frac{x_k}{T_N-x_j}\\
&= \frac{x_k/T_N}{D}\left(\frac{T_N-x_k}{T_N-2x_k}+\sum_{\substack{j\in U\\ j\neq k}}\frac{x_j}{T_N-2x_j}\right)\\
&= \frac{x_k/T_N}{D}\left(\frac{T_N}{T_N-2x_k}-\frac{2x_k}{T_N-2x_k}+\sum_{j\in U}\frac{x_j}{T_N-2x_j}\right)\\
&= \frac{x_k/T_N}{D}\left(1+\sum_{j\in U}\frac{x_j}{T_N-2x_j}\right)=\frac{x_k/T_N}{D}(2D)=\frac{2x_k}{T_N}
\end{aligned}
$$

Donde

$$
\begin{aligned}
D &= \sum_{k\in U}\frac{x_k(T_N-x_k)}{T_N(T_N-2x_k)}\\
&= \frac{1}{2}\sum_{k\in U}\frac{x_k(2T_N-2x_k)}{T_N(T_N-2x_k)}\\
&= \frac{1}{2}\left(1+\sum_{k\in U}\frac{x_k}{T_N-2x_k}\right)
\end{aligned}
$$

La última relación se tiene puesto que

$$
\sum_{k\in U}\frac{x_k(T_N-x_k)}{T_N(T_N-2x_k)}-\sum_{k\in U}\frac{x_k}{T_N-2x_k}=1
$$

Análogamente para las probabilidades de inclusión de segundo orden. ■

Resultado 4.4.2. *Bajo muestreo πPT con el algoritmo de selección de Brewer, se tiene que.*

1. *$Var_{\pi PT}(\hat{t}_{y,\pi})$ es menor que $Var_{PPT}(\hat{t}_{y,p})$.*
2. *La estimación de la varianza es siempre positiva.*

Lohr (2000) afirma que generalmente el muestreo con reemplazo es menos eficiente que el muestreo sin reemplazo. Sin embargo, el muestreo con reemplazo se

utiliza con mucha más frecuencia debido a la facilidad que brinda para elegir y analizar las muestras. Se ha investigado mucho acerca del muestreo con probabilidades proporcionales sin reemplazo; hay que notar que la teoría de éstos tipos de muestreo es mucho más complicada. Existen varios algoritmos que permiten la selección de muestras de tamaño $n > 2$ con probabilidades de inclusión desiguales; en particular, con probabilidades proporcionales a una característica de información auxiliar[3]. En esta sección, revisaremos algunos de estos esquemas que permiten la selección de muestras para tamaños de muestra fijos y mayores que dos.

4.4.1 Método de Sunter

En Sunter (1977) y en Sunter (1986) se propune un procedimiento secuencial que, en general, no es aplicable a cualquier vector de probabilidades de inclusión de primer orden. Este algoritmo de muestreo sólo funciona cuando los elementos de la población son ordenados descendentemente y cuando los elementos con valores más pequeños comparten las mismas probabilidades de inclusión. Este método, que en realidad es una modificación del algoritmo de Fan-Muller-Rezucha para la selección de muestras simples, asume la existencia de una variable auxiliar que induce probabilidades de inclusión de primer orden dadas por la expresión (4.3.1) y consiste en:

1. Ordenar descendentemente la población de acuerdo con los valores que toma la característica de información auxiliar x_k.
2. Realizar $\xi_k \sim U(0,1)$.
3. Para $k = 1$, el primer elemento de la lista ordenada es incluido en la muestra sí y solamente sí $\xi_1 < \pi_1$.
4. Para $k \geq 2$, el k-ésimo elemento de la lista ordenada es incluido en la muestra sí y solamente sí
$$\xi_k \leq \frac{n - n_{k-1}}{n - \sum_{i=1}^{k-1} \pi_i} \pi_k$$
donde n_{k-1} representa el número de elementos que ya han sido seleccionados al final del paso $k-1$.

Resultado 4.4.3. *Bajo el esquema de selección de Sunter, las probabilidades de inclusión de primer orden están dadas por*

$$\pi_k = \begin{cases} \dfrac{nx_k}{T_N} & \text{si } k = 1, \ldots, k^* - 1 \\ \\ \dfrac{n\bar{x}_{k^*}}{T_N} & \text{si } k = k^*, \ldots, N \end{cases}$$

donde $k^* = \min\{k_0, N - n + 1\}$ *con* k_0 *equivalente al menor* k *para el cual se cumple que* $nx_k/T_k > 1$, $T_k = \sum_{j=1}^{k} x_j$ *y*

[3] El lector interesado en conocer aún más acerca de estos algoritmos de selección puede referirse a los siguientes tres libros: Brewer & Hanif (1983), Hájek (1981) y Tillé (2006)).

$$\bar{x}_{k^*} = \frac{T_{k^*}}{N - k^* + 1}$$

Por otra parte, se cumple que para todo $k \neq l$, $\pi_{kl} > 0$ y $\Delta_{kl} < 0$.

Con el anterior resultado se establece que este método de selección de muestras no induce probabilidades de inclusión estrictamente proporcionales a la característica de información auxiliar. Särndal, Swensson & Wretman (1992) afirman que relajar un poco este supuesto es un precio menor que debe pagarse para que el esquema de selección sea ejecutable en la práctica.

Ejemplo 4.4.1. Volviendo con la población ejemplo U. Suponga que se tiene acceso a los valores de la característica de información auxiliar x para todos los elementos de la población. Es posible seleccionar una muestra πPT de tamaño $n = 3$ con el método de Sunter. Para tal fin, es necesario recurrir a la función `S.piPS` del paquete `TeachingSampling`.

Esta función consta de tres argumentos: el primero, `x`, hace referencia al vector de información auxiliar continua para toda la población. El segundo, `n`, determina el tamaño de la muestra. Con estos dos argumentos, la función `S.piPS` construye las probabilidades de inclusión proporcionales a la característica de información auxiliar. El tercer argumento, `e`, que es opcional, corresponde a un vector de números aleatorios con el que se procede a ejecutar el esquema de selección de Sunter.

```
> U <- c("Yves", "Ken", "Erik", "Sharon", "Leslie")
> N <- length(U)
> n <- 3
> x <- c(52,60,75,100,50)
> pi <- (n*x)/sum(x)
> pi
[1] 0.46 0.53 0.67 0.90 0.44
> sum(pi)
[1] 3

> sam <- S.piPS(n,x,e=runif(N))
> U[sam]
[1] "Sharon" "Erik"   "Ken"
> x[sam]
[1] 100  75  60
```

La función `S.piPS` devuelve un conjunto de índices (distintos por definición) que aplicados a los rótulos poblacionales proporcionan la muestra realizada o seleccionada. Para el anterior ejercicio particular, la muestra realizada estuvo conformada por **Sharon**, **Erik** y **Ken**. Es importante recalcar que esta función no necesita de ningún ordenamiento previo sobre la característica de información auxiliar; en otras palabras, los resultados serán idénticos si se realiza un ordenamiento previo o si no se realiza tal ordenamiento.

4.4.2 Método de escisión

Desde la publicación de Brewer & Hanif (1983) se han propuesto numerosas técnicas de muestreo con probabilidades de inclusión desiguales. Sin embargo, en el artículo de Deville & Tillé (1998), se habla de ocho nuevos métodos; entre ellos, el método de escisión. Este método es considerado como un nuevo enfoque que presenta de manera más simple los restantes métodos de selección de muestras con probabilidades desiguales. Tillé (2006) comenta que el método de escisión es un medio para integrar la presentación de los demás métodos y para hacerlos comparables.

En palabras de uno de los autores (Tillé 2006), el método de escisión propuesto por Deville & Tillé (1998) es:

> ...un marco de referencia de los métodos de muestreo sin reemplazo, con tamaño muestral fijo y con probabilidades desiguales, en particular con probabilidades proporcionales al tamaño de una característica de información auxiliar.
>
> La idea básica del método consiste en dividir el vector de probabilidades de inclusión en dos o más vectores nuevos. A continuación, uno de estos vectores se selecciona aleatoriamente, de tal manera que el promedio de los vectores de como resultado el vector de probabilidades de inclusión. Este simple paso se repite hasta que se obtenga una muestra.
>
> Con el planteamiento anterior, el método de escisión se puede considerar como un algoritmo de Martingalas que incluye todos los procedimientos de selección individual y secuencial y permite derivar un gran número de algoritmos de muestreo de probabilidades desiguales. Más aun, muchos procedimiento bien conocidos de probabilidades desiguales pueden ser formulados bajo la forma de una partición del vector de probabilidades de inclusión. Por tanto, la presentación puede ser estandarizada, lo cual permite una comparación más simple de procedimientos.

Escisión en dos partes

Este método consiste en seleccionar una muestra, de tamaño $n(S) = n$, de probabilidades desiguales mediante la partición de la probabilidad de inclusión del k-ésimo elemento en dos partes π_k^a y π_k^b tal que

$$\pi_k = \lambda\pi_k^a + (1-\lambda)\pi_k^b \tag{4.4.1}$$

De tal forma que $0 \leq \pi_k^a \leq$ y $0 \leq \pi_k^b \leq$ y que

$$\sum_{k\in U}\pi_k^a = \sum_{k\in U}\pi_k^b = n \tag{4.4.2}$$

Donde $0 < \lambda < 1$. La esencia del método es la selección de n elementos con probabilidades desiguales mediante la transformación iterativa del vector de probabilidades de inclusión. Si la escisión es tal que uno o varios de los π_k^a y de los π_k^b son

equivalentes a cero o uno, entonces el problema de muestreo se verá reducido en el siguiente paso. De hecho, un vez que un componente del vector de probabilidades de inclusión converja a cero o uno, es deberá permanecer en este estado hasta que se seleccione una muestra[4]. En general, el algoritmo de muestreo de este esquema es el siguiente:

1. Definir $\boldsymbol{\pi}(0) = \boldsymbol{\pi}$.
2. Construir un par de vectores $\boldsymbol{\pi}^a(t)$ y $\boldsymbol{\pi}^b(t)$ y definir un número $\lambda(t) \in (0,1)$ tales que
$$\boldsymbol{\pi}(t) = \lambda(t)\boldsymbol{\pi}^a(t) + (1-\lambda(t))\boldsymbol{\pi}^b(t) \tag{4.4.3}$$
3. Definir para el siguiente paso al vector de probabilidades de inclusión de tal forma que
$$\boldsymbol{\pi}(t+1) = \begin{cases} \boldsymbol{\pi}^a(t) & \text{con probabilidad } \lambda(t) \\ \boldsymbol{\pi}^b(t) & \text{con probabilidad } 1-\lambda(t) \end{cases} \tag{4.4.4}$$
4. Iterar hasta obtener convergencia; es decir, hasta que todas las entradas del vector de probabilidades de inclusión sean cero o uno en ambas particiones. De esta forma, para cada tiempo t, existe una posible muestra correspondiente a $S = \boldsymbol{\pi}(t)$.

Esquema de soporte mínimo

Definición 4.4.1. *Si para un vector fijo de probabilidades de inclusión es posible plantear un diseño de muestreo cuyo soporte contenga a lo más N muestras s, tales que $p(s) > 0$. En tal caso, el diseño de muestreo se dice de soporte mínimo.*

A continuación se presenta el esquema de soporte mínimo que permite seleccionar una muestra en a lo más N pasos.

Paso 1 Ordenar el vector de probabilidades de inclusión en orden ascendente, denotado como $(\pi_{(1)}, \ldots, \pi_{(k)}, \ldots, \pi_{(N)})$

Paso 2 (Primera iteración, $t = 1$) Calcular

$$\lambda(1) = \text{mín}\{1 - \pi_{(N-n)}, \pi_{(N-n+1)}\}$$

Luego, computar las siguientes particiones del vector de probabilidades de inclusión

$$\pi^a_{(k)}(1) = \begin{cases} 0 & \text{si } k \leq N-n \\ 1 & \text{si } k > N-n \end{cases} \tag{4.4.5}$$

$$\pi^b_{(k)}(1) = \begin{cases} \frac{\pi_{(k)}}{1-\lambda(1)} & \text{si } k \leq N-n \\ \\ \frac{\pi_{(k)}-\lambda(1)}{1-\lambda(1)} & \text{si } k > N-n \end{cases} \tag{4.4.6}$$

[4]Una muestra es seleccionada cuando todas las entradas del vector de probabilidades de inclusión se conviertan en ceros o unos.

Paso 3 (t-ésima iteración, $t \geq 2$) Definir los siguientes conjuntos

$$A(t) = \{k|0 < \pi^b_{(k)}(t-1) < 1\}$$
$$B(t) = \{k|\pi^b_{(k)}(t-1) = 1\}$$

y las siguientes cantidades:

$$N^*(t) = \#A(t)$$
$$n^*(t) = \#B(t)$$

Luego, para los elementos $k \in A(t)$ calcular

$$\lambda(t) = \text{mín}\{1 - \pi^b_{(N^*(t)-n^*(t))}(t-1), \pi^b_{(N^*(t)-n^*(t)+1)}(t-1)\}$$

A continuación, para los elementos $k \in A(t)$ computar las siguientes particiones del vector de probabilidades de inclusión

$$\pi^a_{(k)}(t) = \begin{cases} 0 & \text{si } k \leq N^*(t) - n^*(t) \\ 1 & \text{si } k > N^*(t) - n^*(t) \end{cases} \tag{4.4.7}$$

$$\pi^b_{(k)}(t) = \begin{cases} \frac{\pi^b_{(k)}(t-1)}{1-\lambda(t)} & \text{si } k \leq N^*(t) - n^*(t) \\ \\ \frac{\pi^b_{(k)}(t-1)-\lambda(t)}{1-\lambda(t)} & \text{si } k > N^*(t) - n^*(t) \end{cases} \tag{4.4.8}$$

Paso 4 Iterar hasta obtener convergencia; es decir, hasta que $\pi^b_{(k)}(t) \in \{0, 1\}$.

Ejemplo 4.4.2. En este apartado se muestra paso a paso cómo trabaja el algoritmo de mínimo soporte basado en el método de escisión. Volvemos entonces a nuestra población ejemplo

$$U = \{\textbf{Yves, Ken, Erik, Sharon, Leslie}\}$$

El cálculo de las probabilidades de inclusión se hace con respecto a la expresión (4.3.1) donde la característica de información auxiliar corresponde a

$$\mathbf{x} = (52, 60, 75, 100, 50)$$

Por tanto, el vector de probabilidades de inclusión está dado por

$$\boldsymbol{\pi} = (0.46, 0.53, 0.67, 0.90, 0.44)$$

El método exige el ordenamiento del vector de probabilidades de inclusión en orden ascendente. Luego de esto, se tiene que el procedimiento converge en cuatro etapas. La tabla 4.3 muestra la convergencia del método y todas las posibles muestras que surgen del diseño muestral con soporte mínimo. Los cálculos en cada etapa se dan a continuación:

Tabla 4.3: *Diseño de mínimo soporte para la población* U.

		Etapa 1 $\lambda(1) = 0.53$		Etapa 2 $\lambda(2) = 0.06$		Etapa 3 $\lambda(3) = 0.02$		Etapa 4 $\lambda(4) = 0.78$	
k	π_k	π_k^a	π_k^b	π_k^a	π_k^b	π_k^a	π_k^b	π_k^a	π_k^b
Leslie	0.44	0	0.94	0	1	1	1	1	1
Yves	0.46	0	0.98	1	0.98	0	1	1	1
Ken	0.53	1	0	0	0	0	0	0	0
Erik	0.67	1	0.29	1	0.24	1	0.22	0	1
Sharon	0.90	1	0.79	1	0.78	1	0.78	1	0

Etapa 1 $N = 5$, $n = 3$, $\lambda = \text{mín}\{1 - \pi_{(2)}, \pi_{(3)}\} = 0.53$

Etapa 2 $N^*(2) = 4$, $n^*(2) = 3$, $\lambda(2) = \text{mín}\{1 - \pi_{(1)}(1), \pi_{(2)}(1)\} = 0.06$

Etapa 3 $N^*(3) = 3$, $n^*(3) = 2$, $\lambda(3) = \text{mín}\{1 - \pi_{(1)}(2), \pi_{(2)}(2)\} = 0.02$

Etapa 4 $N^*(4) = 2$, $n^*(4) = 1$, $\lambda(4) = \text{mín}\{1 - \pi_{(1)}(3), \pi_{(2)}(3)\} = 0.78$

Por tanto, el diseño muestral de mínimo soporte está dado por

$$p(s) = \begin{cases} 0.53 & \text{si } s = \{\textbf{Ken, Erik, Sharon}\} \\ 0.0282 = (1 - 0.53) \times 0.06 & \text{si } s = \{\textbf{Yves, Erik, Sharon}\} \\ 0.0088 = (1 - 0.53 - 0.0282) \times 0.02 & \text{si } s = \{\textbf{Leslie, Erik, Sharon}\} \\ 0.3377 = (1 - 0.53 - 0.0282 - 0.008) \times 0.78 & \text{si } s = \{\textbf{Leslie, Yves, Sharon}\} \\ 0.0953 = (1 - 0.53 - 0.0282 - 0.008 - 0.3377) & \text{si } s = \{\textbf{Leslie, Yves, Erik}\} \end{cases}$$

4.4.3 Estimación de la varianza

Existe un número muy grande de diseños y algoritmos de muestreo que trabajan bajo el supuesto de probabilidades de inclusión desiguales. En el caso particular del diseño de muestreo sin reemplazo y proporcional al tamaño de una característica de interés, las probabilidades de inclusión siguen el comportamiento dado por la expresión (4.3.1). Cada uno de estos métodos de muestreo inducen probabilidades de inclusión de primer y segundo orden. Las probabilidades de inclusión de primer orden son esenciales al momento de completar la estrategia de muestreo con el estimador de Horvitz-Thompson. Sin embargo, las probabilidades de inclusión de segundo orden, aunque servirían teóricamente para calcular y estimar la varianza del estimador de Horvitz-Thompson, son ineficientes pues cuando el tamaño de muestra crece, su cálculo se vuelve una total aventura, en muchos casos imposible de finiquitar.

Al respecto Tillé (2006) comenta, en el prefacio de su libro de algoritmos de muestreo, que «tiene la convicción de que las probabilidades de inclusión de segundo orden no son usadas para nada» y añade que «en la práctica el uso de las probabilidades de inclusión de segundo orden es muchas veces irreal porque son muy difíciles de calcular computacionalmente y n^2 términos deben ser sumados para calcular la estimación».

Para evitar el cálculo y estimación de la varianza del estimador de Horvitz-Thompson con dobles sumas, Deville & Tillé (2005) proponen una aproximación de la varian-

za[5] y su respectiva estimación para un diseño exponencial[6] dada por el siguiente resultado

Resultado 4.4.4. *Para la familia de diseños exponenciales, la aproximación de la varianza del estimador de Horvitz-Thompson está dada por*

$$Var(\hat{t}_{y,\pi}) = \sum_{k\in U} \frac{b_k}{\pi_k^2}(y_k - y_k^*)^2 \tag{4.4.9}$$

donde

$$y_k^* = \pi_k \frac{\sum_{l\in U} b_l y_l/\pi_l}{\sum_{l\in U} b_l} \tag{4.4.10}$$

Hájek (1981) ha propuesto la siguiente escogencia de b_k

$$b_k = \frac{N\pi_k(1-\pi_k)}{(N-1)} \tag{4.4.11}$$

Un estimador de la anterior aproximación de la varianza está dada por

$$\widehat{Var}(\hat{t}_{y,\pi}) = \sum_{k\in S} \frac{c_k}{\pi_k^2}(y_k - \hat{y}_k^*)^2 \tag{4.4.12}$$

donde

$$\hat{y}_k^* = \pi_k \frac{\sum_{l\in S} c_l y_l/\pi_k}{\sum_{l\in S} c_l} \tag{4.4.13}$$

Deville (1993) ha propuesto la siguiente escogencia de c_k

$$c_k = (1-\pi_k)\frac{n}{(n-1)} \tag{4.4.14}$$

Ejemplo 4.4.3. Para nuestra población de ejemplo U, existen $\binom{N}{n} = 10$ posibles muestras πPT de tamaño $n = 3$. Utilizando las probabilidades de inclusión del ejemplo 4.4.1, realice el cálculo léxico-gráfico del estimador de Horvitz-Thompson, calcule la aproximación de la varianza dada por la expresión (4.4.9) y para cada muestra estime esta varianza usando la expresión (4.4.12) y compruebe su insesgamiento.

Acerca del muestreo πPT

En general, la familia de diseños de muestreo πPT son utilizados cuando el comportamiento de la característica de interés en la población finita es bastante asimétrico. Para la estimación de totales, este diseño es más eficiente, en términos de reducción de la varianza. Sin embargo, cuando se quiere estimar otro tipo de parámetros poblacionales, como razones o medianas, los diseños de muestreo proporcionales al tamaño no son muy apetecidos, pues es difícil encontrar una característica de información auxiliar bien correlacionada con la razón entre las dos características de interés. En resumen, se tiene que:

[5] Existe mucha literatura escrita alrededor del tema de aproximaciones y simplificaciones de la varianza del estimador de Horvitz-Thompson. Para una mejor comprensión del tema Matei & Tille (2005) han escrito un excelente artículo de revisión.

[6] Los diseños de muestreo exponenciales son una gran familia que incluyen diseños tales como muestreo aleatorio simple, muestreo multinomial, muestreo de probabilidades desiguales con reemplazo y algunos diseños de probabilidades desiguales sin reemplazo. Para más información acerca de los diseños de muestreo exponenciales el lector deberá remitirse a Tillé (2006).

- Se utiliza esencialmente para la estimación de totales poblacionales.
- En encuestas de hogares no vale la pena utilizar este diseño pues, en general, en cada vivienda hay una misma cantidad de hogares.
- En encuestas de negocios es bueno utilizar diseños proporcionales porque sí existen diferencias en los tamaños considerados.
- Debido a que este diseño de muestreo involucra información auxiliar, entonces es más eficiente que el diseño de muestreo aleatorio simple.
- Un defecto de este diseño de muestreo es que su varianza no es una función monótona decreciente. Debido a la configuración particular de la información, la varianza puede crecer si se aumenta el tamaño de muestra.

4.4.4 Marco y Lucy

En este apartado de Marco y Lucy suponga que se tienen las mismas condiciones que en el apartado de Marco y Lucy del diseño de muestreo PPT (ver la sección 4.2.4). Siendo así, el marco de muestreo permite conocer los valores poblacionales de una característica de información auxiliar. En este caso ésta es la variable `Income`. Dadas las bondades del marco de muestreo, se quiere seleccionar una muestra de tamaño `n=400` mediante un diseño de muestreo sin reemplazo que induzca probabilidades de inclusión proporcionales a esta característica de información auxiliar.

La selección de la muestra se realiza haciendo uso de la función `S.piPS` del paquete `TeachingSampling` para la cual los argumentos introducidos son: el vector de valores poblacionales de la característica de información auxiliar `Income` y el tamaño de la muestra sin reemplazo `n=400`. Nótese que esta función utiliza el algoritmo de selección de Sunter.

```
> data(Lucy)
> attach(Lucy)
> N <- dim(Lucy)[1]
> n <- 400
> pik <- n*Income/sum(Income)
> sam <- S.piPS(n, Income)
> muestra <- Lucy[sam,]
> attach(muestra)
> muestra

     Identificador Ubicación   Level Zona Income
2338        AB1132    c25k61  Grande    A   2510
2370         AB749    c25k93  Grande    D   1911
2329        AB1096    c25k52  Grande    A   1620
...
1367         AB388    c14k80 Pequeño    C     96
284          AB148     c3k86 Pequeño    B     91
755          AB191     c8k62 Pequeño    C     76
```

El resultado de la función `S.piPS` es una muestra ordenada de forma descendente por los valores de la característica de información auxiliar. El siguiente paso es recolectar la información de las características de interés `Employees` e `Taxes` para los elementos incluidos en la muestra realizada.

Después de recolectar la información, es necesario estimar los totales de las características de interés. En esta etapa se utiliza la función `E.piPS` del paquete `TeachingSampling` cuyos argumentos son: `estima`, correspondiente a la lista que contiene los valores observados en la muestra para cada una de las características de interés y `pik.s`, correspondiente al vector de probabilidades de inclusión (proporcionales a la característica de información auxiliar) de los elementos en la muestra.

```
> pik.s <- pik[sam]
> estima <- data.frame(Income, Employees, Taxes)

> E.piPS(estima, pik.s)
                                Income    Employees        Taxes
Total estimado            1.035217e+06 1.515538e+05 2.821118e+04
Varianza                  3.559408e-23 2.288151e+07 4.403830e+05
coeficiente de variación 5.763117e-16 3.156279e+00 2.352307e+00
```

Los resultados para este ejercicio particular son excelentes. Nótese que los estimativos de la varianza no son exactos, pues están dados por el resultado 4.4.2, aunque sí aproximados. Por otra parte, el resultado 4.3.4 asegura que éstos serían menores a los arrojados por la estrategia de muestreo que utiliza un diseño PPT con reemplazo y el estimador de Hansen-Hurwitz. Por supuesto, este diseño de muestreo es más eficiente que el de Poisson, no es de extrañar que los resultados para la variable Ingreso sean tan exactos. Recuérdese que ésta fue la variable utilizada como característica de información auxiliar. La siguiente tabla muestra los resultado para un ejercicio particular. Una vez más, la característica Impuestos tiene un menor coeficiente de variación estimado puesto que está mucho mejor correlacionada con la variable Ingreso.

Véase que para obtener estos resultados, fue necesario conocer el valor de N dado por la longitud del vector de información auxiliar. Nótese que no siempre se puede asegurar el conocimiento del total poblacional. Sin embargo, aunque no se conociera, con la función `HT` se hubiera llegado a los mismos resultados, en términos de la estimación de los totales, pero no se obtendrían los estimativos concernientes a la varianza, tal y como se ilustra a continuación.

```
> HT(estima, pik.s)
     Income     Employees      Taxes
[1,] 1035217     151553.8   28211.18
```

4.5 Ejercicios

4.1 Demuestre o refute la siguiente afirmación: ≪Cuando el comportamiento de la característica de interés es uniforme en la población es más conveniente

Tabla 4.4: *Muestreo πPT: estimación de los totales de las características de interés.*

Variable	Total poblacional	Total estimado	cve %	Desv. %
Ingreso	1035217	1035217	0.000	0.000
Empleados	151950	151553	3.156	-0.002
Impuestos	28654	28211	2.352	-0.015

utilizar diseños de muestreo proporcionales al tamaño de una característica de información auxiliar≫.

4.2 Demuestre o refute la siguiente afirmación: ≪En muestreo Poisson, cuando las probabilidades de inclusión son tales que $\pi_k = ny_k/t_y$ la varianza del estimador de Horvitz-Thompson es nula≫.

4.3 Complete el cálculo léxico-gráfico del ejemplo 4.1.2.

4.4 Suponga una población de 10 elementos $U = \{e_1, \ldots, e_{10}\}$ cuyo marco de muestreo contiene una característica de información auxiliar dada por

$$\mathbf{x} = (62, 151, 76, 77, 80, 60, 194, 78, 74, 61)$$

- Si se desea seleccionar una muestra sin reemplazo de tamaño esperado $n(S) = 6$, utilice la expresión (4.3.2) para construir un vector de probabilidades de inclusión proporcionales a $\mathbf{x}$ tales que $0 < \pi_k \leq 1$ para todo $k \in U$ y verifique $\sum_U \pi_k = 6$
- Utilice el algoritmo de la sección 4.1.1 para seleccionar una muestra Poisson teniendo en cuenta que se obtuvo el siguiente conjunto de números aleatorios uniformes

$$\boldsymbol{\varepsilon} = \{0.858, 0.698, 0.541, 0.320, 0.965, 0.497, 0.208, 0.006, 0.340, 0.206\}$$

- Utilice el método de Sunter para seleccionar una muestra πPT teniendo en cuenta que se obtuvo el siguiente conjunto de números aleatorios uniformes

$$\boldsymbol{\xi} = \{0.322, 0.542, 0.032, 0.141, 0.453, 0.668, 0.174, 0.318, 0.691, 0.006\}$$

4.5 (Särndal, Swensson & Wretman 1992, p. 117) Para estimar el total de la característica de interés y de una población de $N = 284$ elementos, se utilizó un diseño de muestreo Poisson de tamaño de muestra esperado $n(S) = 10$. Las probabilidades de inclusión fueron proporcionales a una característica de información auxiliar x cuyo total poblacional es $t_x = 8182$. Luego, el algoritmo de selección arrojó una muestra de tamaño efectivo de 12 elementos, para las cuales se obtuvo la siguiente información

- Calcule una estimación insesgada para el total poblacional de la característica de interés, reporte el coeficiente de variación estimado y un intervalo de confianza al 95 %.
- Calcule una estimación insesgada para la media poblacional de la característica de interés, reporte el coeficiente de variación estimado y un intervalo de confianza al 95 %.

x_k	y_k
54	5246
671	59877
28	2208
27	2546
29	2903
62	6850
42	3773
48	4055
33	4014
446	38945
12	1162
46	4852

- Utilice el estimador alternativo para calcular estimaciones tanto del total como de la media poblacional.

4.6 Complete el cálculo léxico-gráfico del ejemplo 4.4.3.

4.7 Suponiendo que los datos del ejercicio 4.5 provienen de un diseño de muestreo πPT, calcule una estimación para el total de la característica de interés. Utilizando la aproximación de la varianza dada en (4.4.12), reporte el coeficiente de variación estimado y un intervalo de confianza al 95 %.

4.8 Utilice el esquema de mínimo soporte para especificar un diseño de muestreo πPT de tamaño $n = 3$ para una población de tamaño $N = 6$ cuyo vector de probabilidades de inclusión de primer orden es

$$\boldsymbol{\pi} = (0.07, 0.17, 0.41, 0.61, 0.83, 0.91)'$$

Demuestre que el procedimiento converge en cuatro pasos que inducen cinco muestras y calcule la probabilidad de selección de cada muestra.

4.9 Demuestre o refute la siguiente afirmación: «En muestreo PPT es posible utilizar los estimadores de Horvitz-Thompson y de Hansen-Hurwitz, al comparar las dos estrategias se tiene que las dos aportan la misma precisión pero diferente confiabilidad».

4.10 Complete el cálculo léxico-gráfico del ejemplo 4.2.2.

4.11 Suponga una población de 12 elementos $U = \{e_1, \ldots, e_{12}\}$ cuyo marco de muestreo contiene una característica de información auxiliar dada por

$$\mathbf{x} = (674, 802, 829, 726, 709, 789, 742, 791, 805, 797, 771, 692)$$

- Si se desea seleccionar una muestra con reemplazo de tamaño $m = 6$, construya un vector de probabilidades de selección proporcionales a $\mathbf{x}$ tales que $0 < p_k \leq 1$ para todo $k \in U$ y verifique $\sum_U p_k = 6$
- Utilice el método acumulativo total para seleccionar una muestra PPT teniendo en cuenta que para cada una de las seis extracciones se generaron los siguientes números aleatorios uniformes

$$\boldsymbol{\varepsilon} = \{0.075, 0.397, 0.280, 0.407, 0.982, 0.782\}$$

- Utilice el método de Lahiri para seleccionar una muestra PPT usando sus propios números aleatorios η y l en cada una de las extracciones.

4.12 Demuestre o refute la siguiente afirmación: «Para la estimación de totales, el diseño PPT es preferido sobre el diseño πPT porque permiten agilizar los cálculos computacionales de varianza y coeficiente de variación».

4.13 Demuestre o refute la siguiente afirmación: «Para la estimación de totales, el diseño PPT siempre es más eficiente que el diseño de muestreo aleatorio simple con reemplazo».

4.14 Suponga una población de $N = 12$ elementos cuyos valores observados para la característica de interés son

$$\mathbf{y} = \{50, 53, 44, 45, 53, 31, 35, 45, 34, 44, 52, 52\}$$

y los valores observados para la característica de información auxiliar son

$$\mathbf{x} = \{1005, 1072, 884, 907, 1068, 625, 705, 909, 692, 891, 1046, 1052\}$$

- Calcule la correlación entre `x` e `y2/x`
- Realice un gráfico de dispersión para `y/x` y explique si se puede afirmar que la razón es constante para los elementos de la población.
- Utilice el análisis de regresión simple para estimar el valor del intercepto y decida si este es estadísticamente diferente de cero.
- Para un tamaño de muestra $m = 6$, utilice la expresión (4.2.13) y los anteriores argumentos para justificar o descalificar la escogencia del diseño de muestreo PPT para esta población.

4.15 Asumiendo que los datos del ejercicio 4.5 provienen de un diseño de muestreo PPT, calcule la estimación de Hansen-Hurwitz para el total de la característica de interés, reporte el coeficiente de variación estimado y un intervalo de confianza al 95 %. También calcule la estimación de Horvitz-Thompson para el total de la característica de interés.

Capítulo 5

Muestreo estratificado

> La estratificación es una de las técnicas más difundidas y usadas en muestreo puesto que tiene funcionalidades estadísticas y administrativas que la hacen atractiva: permite tratar con subpoblaciones, aumenta la eficiencia de las estimaciones y contribuye a la administración eficiente de grandes encuestas.
>
> Richard Valliant (2000)

En algunas ocasiones, la característica de interés tiende a tomar distintos valores promedio con respecto a subgrupos poblacionales. De alguna manera, si la población tiene un comportamiento diferente en estos subgrupos, es posible mejorar la precisión de las estimaciones tomando muestras independientes en cada uno de los subgrupos poblacionales. Lo anterior es intuitivo cuando entre los subgrupos existe mucha variabilidad, pero dentro de ellos la variabilidad es constante.

En general, cuando existe en el marco de muestreo información auxiliar que permite la división de la población en H subgrupos con el objetivo de seleccionar una muestra en cada subgrupo, se dice que la estrategia de muestreo utiliza un **diseño de muestreo estratificado** y el nombre de los subgrupos, formados antes de la recolección de la información, se denomina **estratos**. Nótese la diferencia con los subgrupos poblacionales llamados **dominios**, en donde la partición de la población se realiza después de la recolección de la información.

Con frecuencia, tenemos información adicional que nos ayuda a diseñar la estrategia de muestreo. Cuando esta información se refiere a la pertenencia de cada uno de los elementos a un subgrupo, podemos aplicar una estrategia que utilice un diseño de muestreo estratificado. No es solamente la disponibilidad de esta información auxiliar la que nos lleva a utilizar un diseño de muestreo estratificado, además de esto:

1. La variable de interés asume distintos valores promedio en diferentes subpoblaciones.

2. De una u otra forma (proceso logístico y/o de recolección de datos) es mejor estratificar y dividir la población en particiones. Lehtonen & Pahkinen (2003)

afirman que algunas variables típicas de estratificación son de tipo regional (municipio, estado o provincia), demográfico (género o grupo de edad) y socioeconómico (grupo de ingresos). Existen censos, en periodos anteriores que pueden contener esta valiosa información.

La necesidad de estratificar[1] la población surge por una o más de las siguientes razones:

- Por razones administrativas. Existen marcos de muestreo que ya tienen dividida la población en subgrupos formados naturalmente.
- Se desea garantizar que la muestra seleccionada sea representativa con respecto al comportamiento de la población según la información auxiliar. Al seleccionar una muestra aleatoria simple de una población de personas, podría suceder que la muestra seleccionada no incluyera a ningún hombre.
- Se requieren estimativos con alta precisión discriminados para cada subpoblación. Aumentar el tamaño de muestra en los estratos menos representados.
- Menor Coste. Distintos esquemas operativos para diversos estratos. Encuestas por correo para empresas grandes. Menor tamaño de muestras en zonas de tolerancia o zonas de difícil manejo del orden público.
- Reducción de la varianza en la estimación. Personas de distintas edades con distintas presiones sanguíneas (estratificar por grupos de edad). Se reduce la varianza pues los estratos son homogéneos por dentro, pero heterogéneos entre sí.

El objetivo del diseño estratificado es dar un tratamiento particular a cada subgrupo, ya sea por razones económicas, administrativas o logísticas. Es indispensable delimitar bien los subgrupos en la etapa de diseño. Por ejemplo, en un estudio dentro de una universidad, si se quiere averiguar el número de horas que los estudiantes permanecen enfrente de un computador, no es una buena idea (defecto técnico) dividir la población en cursos porque los cursos no brindan una partición de la población, dado que en distintos cursos pueden estar los mismos estudiantes.

5.1 Fundamentos teóricos

Suponga que el marco de muestreo es tal que permite conocer la pertenencia de cada elemento de la población U en H sub-grupos poblacionales separados U_h ($h = 1, 2, \ldots, H$) también llamados estratos. Éstos se definen como grupos de elementos mutuamente excluyentes. Cada elemento puede pertenecer a uno y sólo a un estrato. De tal forma que

- $\bigcup_{h=1}^{H} U_h = U$
- $U_h \bigcap U_i = \emptyset \qquad h \neq i$

[1] Dividir la población en H estratos disjuntos.

Cada estrato U_h es de tamaño N_h, por tanto

$$\sum_{h=1}^{H} N_h = N \tag{5.1.1}$$

Con la población dividida en H estratos, el objetivo sigue siendo estimar los siguientes parámetros poblacionales

1. El total poblacional,

$$t_y = \sum_{k\in U} y_k = \sum_{h=1}^{H}\sum_{k\in U_h} y_k = \sum_{h=1}^{H} t_{yh} \tag{5.1.2}$$

   donde $t_{yh} = \sum_{k\in U_h} y_k$

2. La media poblacional,

$$\bar{y} = \frac{\sum_{k\in U} y_k}{N} = \frac{1}{N}\sum_{h=1}^{H}\sum_{k\in U_h} y_k = \frac{1}{N}\sum_{h=1}^{H} N_h\bar{y}_h \tag{5.1.3}$$

   donde $\bar{y}_h = \dfrac{1}{N_h}\sum_{k\in U_h} y_k$

Sampath (2001) afirma que dependiendo de la naturaleza de los estratos, diferentes estrategias de muestreo pueden ser utilizadas en diferentes estratos. De tal forma que, en ausencia de información auxiliar, se utilice una estrategia aleatoria simple en algunos estratos, mientras que para aquellos sub-grupos tales que el marco de muestreo permita el conocimiento de información auxiliar continua, es posible aplicar una estrategia de muestreo proporcional al tamaño, e incluso para aquellos sub-grupos en los que, por obligación (logística o técnica), se deba aplicar un censo.

Es importante aclarar que la selección de las H muestras es realizada de manera independiente en cada estrato.[2] De tal forma que la muestra aleatoria S[3] queda definida por

$$S = \bigcup_{h=1}^{H} S_h. \tag{5.1.4}$$

En particular, si la muestra seleccionada es s, entonces

$$s = \bigcup_{h=1}^{H} s_h. \tag{5.1.5}$$

[2]Esto se debe a la independencia entre las selecciones. Aunque se conozcan qué unidades serán incluidas en la muestra de algún estrato, este conocimiento no afecta, de ninguna manera, la inclusión de cualquier otra unidad en los restantes estratos.

[3]Nótese que S es una variable aleatoria y que las medidas de probabilidad utilizadas para la selección de muestras en cada estrato son distintas.

Nótese que si el tamaño de muestra en cada estrato es igual a n_h, entonces el tamaño de la muestra seleccionada mediante un diseño de muestreo estratificado es

$$n = \sum_{h=1}^{H} n_h. \tag{5.1.6}$$

Así, para cada estrato $h \quad h = 1, \ldots, H$ existe un conjunto de todas las posibles muestras denotado como soporte del estrato h, o Q_h. Cada uno de los soportes Q_h induce la definición del soporte general de la siguiente manera

$$Q^H = \bigcup_{h=1}^{H} Q_h. \tag{5.1.7}$$

La cardinalidad de cada soporte Q_h depende del diseño de muestreo utilizado en la selección de la muestra del estrato h. Así

$$\#Q^H = \prod_{h=1}^{H} \#Q_h. \tag{5.1.8}$$

Por supuesto, el diseño de muestreo estratificado es un autentico diseño de muestreo como lo enuncian los siguientes resultados.

Resultado 5.1.1. *Siendo $p_1(s_1), p_2(s_2), \ldots, p_H(s_H)$ los diseños de muestreo utilizados en cada estrato $h \quad h = 1, \ldots, H$, entonces el diseño de muestreo estratificado se define como*

$$p(s) = \prod_{h=1}^{H} p_h(s_h) \tag{5.1.9}$$

Prueba. Se tiene que

$$\begin{aligned} p(s) &= Pr(\text{Seleccionar } s_1 \text{ de } U_1, \cdots, \text{Seleccionar } s_H \text{ de } U_H,) \\ &= p_1(s_1) \cdots p_H(s_H), \end{aligned}$$

puesto que el proceso de selección es independiente en cada estrato. ■

Resultado 5.1.2. *El diseño de muestreo estratificado cumple que*

1. *$p(s) \geq 0$ para todo $s \in Q$*
2. *$\sum_{s \in Q} p(s) = 1$*

Prueba. La primera propiedad se tiene de inmediato puesto que todas las expresiones en 5.1.9 son mayores o iguales a cero. La segunda propiedad se tiene por inducción matemática sobre el número de estratos.

- Si $H = 2$ existen dos soporte, uno para cada estrato, Q_1 definido como

$$Q_1 = \{s_{11}, s_{12}, \ldots, s_{1H_1}\} \tag{5.1.10}$$

y Q_2 definido como

$$Q_2 = \{s_{21}, s_{22}, \ldots, s_{2H_2}\} \tag{5.1.11}$$

tales que

$$Q^2 = \left\{ s_{11} \bigcup s_{21}, s_{11} \bigcup s_{22}, \ldots, s_{11} \bigcup s_{2H_2}, \ldots, s_{1H_1} \bigcup s_{2H_2} \right\} \tag{5.1.12}$$

Ahora, como la selección de las muestras se realiza en forma independiente, en particular se tiene que

$$p\left(s_{11} \bigcup s_{21}\right) = p(s_{11})p(s_{21}) \tag{5.1.13}$$

de manera análoga para el elemento que pertenezca al soporte. Ahora,

$$\begin{aligned}
\sum_{s \in Q} p(s) &= p(s_{11})p(s_{21}) + p(s_{11})p(s_{22}) + \ldots + p(s_{11})p(s_{2H_2}) + \\
&\ldots + p(s_{1H_1})p(s_{21}) + p(s_{1H_1})p(s_{22}) + \ldots + p(s_{1H_1})p(s_{2H_2}) \\
&= p(s_{11})[\underbrace{p(s_{21}) + p(s_{22}) + \ldots + p(s_{2H_2})}_{1}] + \\
&\ldots + p(s_{1H_1})[\underbrace{p(s_{21}) + p(s_{22}) + \ldots + p(s_{2H_2})}_{1}] \\
&= p(s_{11}) + \ldots + p(s_{1H_1}) \\
&= 1
\end{aligned}$$

- Si $H = k$, se supone que

$$\sum_{s \in Q^k} p(s) = 1 \tag{5.1.14}$$

donde

$$Q^k = \left\{ \bigcup_{h=1}^{k} s_h \quad | \quad s_h \in Q_h \right\}. \tag{5.1.15}$$

- Si $H = k + 1$, se tienen $k + 1$ soportes tales que

$$\begin{aligned}
Q_1 &= \{s_{11}, s_{12}, \ldots, s_{1H_1}\} \\
&\vdots \\
Q_k &= \{s_{k1}, s_{k2}, \ldots, s_{kH_k}\} \\
Q_{k+1} &= \{s_{k+1,1}, s_{k+1,2}, \ldots, s_{k+1,H_{k+1}}\}
\end{aligned} \tag{5.1.16}$$

Por consiguiente se tiene que

$$\begin{aligned}
\sum_{s \in Q} p(s) &= p(s_{k+1,1}) \left[\underbrace{\sum_{s \in Q^k} p(s)}_{1} \right] + \ldots + p(s_{k+1,1H_{k+1}}) \left[\underbrace{\sum_{s \in Q^k} p(s)}_{1} \right] \\
&= p(s_{k+1,1}) + \ldots + p(s_{k+1,H_{k+1}}) \\
&= 1
\end{aligned}$$



5.1.1 Estimación en el muestreo estratificado

Si uno de los propósitos de la estratificación es obtener estimaciones más precisas, cabe preguntarse qué forma toman los estimadores y cómo definirlos a través de los estratos; pero aun más ¿qué forma toma la varianza del estimador en los estratos y su varianza estimada?. Los siguientes resultados, responden a los anteriores cuestionamientos.

Resultado 5.1.3. *Si $\hat{t}_{yh}$ estima insesgadamente el total de la característica de interés t_{yh} del subgrupo poblacional h con varianza igual a $Var(\hat{t}_{yh})$, entonces un estimador insesgado para el total poblacional t_y está dado por*

$$\hat{t}_y = \sum_{h=1}^{H} \hat{t}_{yh} \tag{5.1.17}$$

el cual tiene una varianza igual a

$$Var(\hat{t}_y) = \sum_{h=1}^{H} Var(\hat{t}_{yh}) \tag{5.1.18}$$

Prueba. Dado que $\hat{t}_{yh}$ es insesgado, tenemos que

$$\begin{aligned} E\left(\sum_{h=1}^{H} \hat{t}_{yh}\right) &= \sum_{h=1}^{H} E\left(\hat{t}_{yh}\right) \\ &= \sum_{h=1}^{H} t_{yh} = t_y \end{aligned}$$

Por otro lado, acudiendo a la independencia de la selección de muestras en cada estrato

$$\begin{aligned} Var\left(\sum_{h=1}^{H} \hat{t}_{yh}\right) &= \sum_{h=1}^{H} Var\left(\hat{t}_{yh}\right) + \sum_{h=1}^{H}\sum_{i=1}^{H} \underbrace{Cov\left(\hat{t}_{yh}, \hat{t}_{yi}\right)}_{0} \\ &= \sum_{h=1}^{H} Var(\hat{t}_{yh}) \end{aligned}$$

■

Resultado 5.1.4. *Si $\widehat{Var}(\hat{t}_{yh})$ estima insesgadamente a $Var(\hat{t}_{yh})$, entonces un estimador insesgado para $Var(\hat{t}_y)$ está dado por*

$$\widehat{Var}(\hat{t}_y) = \sum_{h=1}^{H} \widehat{Var}(\hat{t}_{yh}) \tag{5.1.19}$$

Prueba. La demostración es inmediata por el insesgamiento en cada uno de los estratos. ■

5.1.2 El estimador de Horvitz-Thompson

Resultado 5.1.5. *Para el diseño de muestreo estratificado, el estimador de Horvitz-Thompson, su varianza y su varianza estimada están dados por:*

$$\hat{t}_{y,\pi} = \sum_{h=1}^{H} \hat{t}_{yh,\pi} \tag{5.1.20}$$

$$Var_{EST}(\hat{t}_{y,\pi}) = \sum_{h=1}^{H} Var_{p_h}(\hat{t}_{yh,\pi}) \tag{5.1.21}$$

$$\widehat{Var}_{EST}(\hat{t}_{y,\pi}) = \sum_{h=1}^{H} \widehat{Var}_{p_h}(\hat{t}_{yh,\pi}) \tag{5.1.22}$$

donde

$$\hat{t}_{yh,\pi} = \sum_{k \in S_h} \frac{y_k}{\pi_k} \tag{5.1.23}$$

Con $Var_{p_e}(\hat{t}_{yh,\pi})$ es la varianza de $\hat{t}_{yh,\pi}$ en el h-ésimo estrato y $\widehat{Var}_{p_h}(\hat{t}_{yh,\pi}$ es la estimación de $Var_{p_h}(\hat{t}_{yh,\pi})$ en el h-ésimo estrato.

Ejemplo 5.1.1. Nuestra población ejemplo U dada por

$$U = \{\textbf{Yves, Ken, Erik, Sharon, Leslie}\}$$

se divide en dos estratos de la siguiente forma

$$U_1 = \{\textbf{Erik, Sharon}\}$$

y el segundo conformado por:

$$U_2 = \{\textbf{Yves, Ken, Leslie.}\}$$

En el primer estrato se selecciona una muestra aleatoria de tamaño $n_1 = 1$ de acuerdo a un diseño de muestreo aleatorio simple sin reemplazo. Por otra parte, en el segundo estrato se selecciona una muestra de tamaño $n_2 = 2$ de acuerdo al siguiente diseño de muestreo

$$p_2(s) = \begin{cases} 1/4, & \text{si } s = \{\textbf{Yves, Ken}\}, \\ 1/4, & \text{si } s = \{\textbf{Yves, Leslie}\}, \\ 1/2, & \text{si } s = \{\textbf{Ken, Leslie}\}. \end{cases}$$

Realice el cálculo léxico-gráfico para comprobar el insesgamiento del estimador de Horvitz-Thompson para todas las posibles muestras de tamaño $n = 3$. Defina los soporte Q_1 y Q_2 así como el soporte general Q^2 para cada estrato.

En las próximas secciones se estudiarán los diseños estratificados más utilizados en la práctica.

5.2 Diseño de muestreo aleatorio estratificado

Al igual que el muestreo aleatorio simple sin reemplazo, el diseño de muestreo aleatorio estratificado (EST-MAS) es el más sencillo de los diseños estratificados. En este caso particular se selecciona una muestra aleatoria simple en cada estrato, de tal forma que las selecciones sean independientes. Este diseño de muestreo es utilizado cuando la variabilidad de la característica de interés dentro de los estratos es similar; en otras palabras, cuando se sabe que el comportamiento de la característica de interés al interior de los estratos es homogéneo. Sin embargo, también se utiliza cuando no se dispone de ninguna información auxiliar continua que permita hacer uso de diseños de muestreo, en cada estrato, que permitan mejorar la eficiencia de una muestra aleatoria simple.

En cada estrato h una muestra aleatoria simple sin reemplazo de tamaño n_h es seleccionada, de manera independiente, de la población del estrato de tamaño N_h. Aunque el diseño de muestreo aleatorio simple es utilizado como un método final de selección de elemento, en conjunto el diseño estratificado puede resultar dramáticamente más eficiente que utilizar un diseño de muestreo aleatorio simple sin dividir la población.

Definición 5.2.1. *Para tamaños de muestra fijos en cada estrato, denotados como $n_1, \ldots, n_H$, un diseño de muestreo se dice estratificado aleatorio simple sin reemplazo si la probabilidad de seleccionar una muestra de tamaño n está dada por*

$$p(s) = \begin{cases} \prod_{h=1}^{H} \frac{1}{\binom{N_h}{n_h}}, & \text{si } \sum_{h=1}^{H} n_h = n \\ 0, & \text{en otro caso} \end{cases} \tag{5.2.1}$$

Nótese que $\sum_{s \in Q^H} p(s) = 1$ porque $\#Q^H = \prod_{h=1}^{H} \binom{N_h}{n_h}$.

5.2.1 Algoritmos de selección

En la selección de las muestras aleatorias simples sin reemplazo en cada estrato es posible utilizar los algoritmos de muestreo dados en el capítulo 3, de tal forma que los siguientes pasos se deben realizar.

- Separar la población en H subgrupos o estratos mediante la caracterización poblacional de información auxiliar.
- En cada estrato seleccionar una muestra aleatoria simple sin reemplazo. Los algoritmos utilizados en la selección de la muestra dentro de cada estrato pueden ser los métodos coordinado negativo o el método de selección y rechazo de Fan, Muller & Rezucha (1962).
- Cada una de las H selecciones es realizada de manera independiente

Ejemplo 5.2.1. Suponga que nuestra población de ejemplo U está particionada de acuerdo a la sección anterior. Es necesario definir los dos estratos en R, de manera tal que ningún elemento tenga una doble pertenencia a algún estrato.

```
> U1  <-  c("Erik", "Sharon")
> N1  <-  length(U1)
> U2  <-  c("Yves", "Ken", "Leslie")
> N2  <-  length(U2)
```

R permite realizar operaciones entre conjuntos de datos. En particular, el operador `union` es utilizado para verificar que la unión de los estratos dé como resultado la población de ejemplo U. Nótese que el tamaño poblacional es la suma de los tamaños de los dos estratos.

```
> U <- union(U1,U2)
> N <-  N1+N2
> U
[1] "Erik"   "Sharon" "Yves"   "Ken"    "Leslie"
> N
[1] 5
```

Se ha decidido seleccionar una muestra aleatoria simple sin reemplazo de tamaño $n_1 = 1$ para U_1 y una muestra aleatoria simple sin reemplazo de tamaño $n_2 = 2$ para U_2. De tal forma que la muestra general será de tamaño $n = n_1 + n_2 = 3$.

```
> sam1 <- sample(N1, 1, replace=FALSE)
> U1[sam1]
[1] "Erik"
>
> sam2 <- S.SI(N2,2)
> U2[sam2]
[1] "Yves" "Ken"
>
> sam <- union(U1[sam1],U2[sam2])
> sam
[1] "Erik" "Yves" "Ken"
```

Por supuesto, es posible utilizar la función `sample` que viene incorporada en el ambiente genérico de R o también es posible utilizar la función la función `S.SI` del paquete `TeachingSampling`. Sin importar el algoritmo de selección de las muestras aleatorias simples sin reemplazo, es importante notar que se han seleccionado tantas muestras como estratos existen en la población.

5.2.2 El estimador de Horvitz-Thompson

La estrategia de muestreo queda definida con el uso del estimador de Horvitz-Thompson. Esta estrategia es la más conocida, aplicada y discutida en los libros de texto. Para esto, el siguiente resultado muestra la construcción de las probabilidades de inclusión.

Resultado 5.2.1. *Para un diseño de muestreo aleatorio estratificado, las probabilidades de inclusión de primer y segundo orden están dadas por:*

$$\pi_k = \frac{n_h}{N_h} \quad \text{si } k \in U_h \tag{5.2.2}$$

$$\pi_{kl} = \begin{cases} \dfrac{n_h}{N_h}, & si\ k = l, k \in U_h, \\ \\ \dfrac{n_h}{N_h}\dfrac{n_h - 1}{N_h - 1}, & si\ k, l \in U_h, \\ \\ \dfrac{n_h}{N_h}\dfrac{n_i}{N_i}, & si\ k \in U_h, l \in U_i, i \neq h. \end{cases} \tag{5.2.3}$$

respectivamente. La covarianza de las variables indicadoras está dada por

$$\Delta_{kl} = \begin{cases} \dfrac{n_h}{N_h}\dfrac{N_h - n_h}{N_h}, & si\ k = l, k \in U_h, \\ \\ -\dfrac{n_h}{N_h^2}\dfrac{(N_h - n_h)}{(N_h - 1)}, & si\ k, l \in U_h, \\ \\ 0, & si\ k \in U_h, l \in U_i, i \neq h. \end{cases} \tag{5.2.4}$$

Prueba. Sea $k \in U_h$

$$\begin{aligned} \pi_k = Pr(k \in S) &= Pr(k \in S_h) \\ &= Pr(I_k(S_h) = 1) \\ &= \frac{\binom{1}{1}\binom{N_h-1}{n_h-1}}{\binom{N_h}{n_h}} = \frac{n_h}{N_h} \end{aligned}$$

por otro lado, si $k, l \in U_h$

$$\begin{aligned} \pi_{kl} &= Pr(k \in S_h \text{ y } l \in S_h) \\ &= Pr(I_k(S_h) = 1 | I_l(S_h) = 1) Pr(I_l(S_h) = 1) \\ &= \frac{n_h - 1}{N_h - 1}\frac{n_h}{N_h} = \frac{n_h}{N_h}\frac{n_h - 1}{N_h - 1} \end{aligned}$$

Pero, si $k \in U_h, l \in U_i, i \neq h$, por la selección independiente en los estrato h e i, se tiene que

$$\begin{aligned} \pi_{kl} &= Pr(k \in S_h \text{ y } l \in S_i) \\ &= Pr(k \in S_h) Pr(l \in S_i) \\ &= \frac{n_h}{N_h}\frac{n_i}{N_i} \end{aligned}$$

■

Una de las razones por las que se utiliza el diseño de muestreo estratificado es porque se desean estimativos de gran precisión en lo subgrupos. Siendo así, al aplicar un diseño EST-MAS se tiene el siguiente resultado que permite obtener estimaciones insesgadas y precisas para cada subgrupo poblacional.

Resultado 5.2.2. *Bajo un diseño de muestreo aleatorio simple sin reemplazo en el estrato h, un estimador insesgado del total t_{yh}, su varianza y su varianza estimada están dados por*

$$\hat{t}_{yh,\pi} = \frac{N_h}{n_h} \sum_{k \in S_h} y_k \tag{5.2.5}$$

$$Var_{MAS}(\hat{t}_{yh,\pi}) = \frac{N_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right) S^2_{y_{U_h}} \tag{5.2.6}$$

$$\widehat{Var}_{MAS}(\hat{t}_{yh,\pi}) = \frac{N_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right) S^2_{y_{S_h}} \tag{5.2.7}$$

respectivamente. En donde

$$S^2_{y_{U_h}} = \frac{1}{N_h - 1} \sum_{k \in U_h} (y_k - \bar{y}_{U_h}), \qquad h = 1, \ldots, H. \tag{5.2.8}$$

la ***varianza poblacional*** *de la característica de interés en el estrato U_h y con*

$$S^2_{y_{S_h}} = \frac{1}{n_h - 1} \sum_{k \in S_h} (y_k - \bar{y}_{S_h}), \qquad h = 1, \ldots, H. \tag{5.2.9}$$

la ***varianza muestral*** *de los valores de la característica de interés en la muestra aleatoria del estrato S_h. Nótese que $\hat{t}_{yh,\pi}$ es insesgado para el total t_{yh} de la característica de interés y, y que $\widehat{Var}_{MAS}(\hat{t}_{yh,\pi})$ es insesgado para $Var_{MAS}(\hat{t}_{yh,\pi})$*

Prueba. Al notar que el subgrupo U_h puede ser tratado como una población separada, la demostración es inmediata al seguir los lineamentos de la demostración del resultado 3.2.4. ■

Una vez se tienen las estimaciones para los subgrupos poblacionales o estratos, se sigue que el total poblacional t_y puede ser estimado usando el siguiente resultado.

Resultado 5.2.3. *Para un diseño de muestreo aleatorio estratificado, el estimador de Horvitz-Thompson del total poblacional t_y, su varianza y su varianza estimada están dados por:*

$$\hat{t}_{y,\pi} = \sum_{h=1}^{H} \hat{t}_{yh,\pi} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{k \in S_h} y_k \tag{5.2.10}$$

$$Var_{MAE}(\hat{t}_{y,\pi}) = \sum_{h=1}^{H} \frac{N_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right) S^2_{yU_h} \tag{5.2.11}$$

$$\widehat{Var}_{MAE}(\hat{t}_{y,\pi}) = \sum_{h=1}^{H} \frac{N_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right) S^2_{ys_h} \tag{5.2.12}$$

respectivamente. Nótese que $\hat{t}_{y,\pi}$ es insesgado para el total t_y de la característica de interés y, y que $\widehat{Var}_{MES}(\hat{t}_{y,\pi})$ es insesgado para $Var_{MAE}(\hat{t}_{y,\pi})$.

Prueba. Dado que $\hat{t}_{yh,\pi}$ estima insesgadamente el total t_{yh} del subgrupo poblacional h con varianza dada por $\frac{N_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right) S^2_{yU_h}$, entonces al utilizar los resultados 5.1.3. y 5.1.4 se tiene de manera inmediata la demostración. ■

Ejemplo 5.2.2. Para nuestra población de ejemplo U, existen $\binom{3}{2}\binom{2}{1} = 6$ posibles muestras de tamaño $n = 3$. Realice el cálculo léxico-gráfico del estimador de Horvitz-Thompson y compruebe el insesgamiento y la varianza.

5.2.3 Estimación de la media poblacional

Una de las formas de conocer si existen diferencias con respecto a los valores que toma la característica de interés en los diferentes estratos, es estimar la media $\bar{y}_{Uh}$ en el subgrupo U_h. De hecho, el diseño estratificado adquiere más validez y ganancia en precisión cuando el comportamiento promedio de la característica de interés es diferente en cada estrato.

Resultado 5.2.4. *Bajo un diseño de muestreo aleatorio simple sin reemplazo en el estrato h, un estimador insesgado de la media $\bar{y}_{Uh}$, su varianza y su varianza estimada están dados por*

$$\hat{\bar{y}}_{Uh,\pi} = \frac{1}{n_h}\sum_{k\in S_h} y_k \tag{5.2.13}$$

$$Var_{MAS}(\hat{\bar{y}}_{Uh,\pi}) = \frac{1}{n_h}\left(1-\frac{n_h}{N_h}\right)S^2_{yU_h} \tag{5.2.14}$$

$$\widehat{Var}_{MAS}(\hat{\bar{y}}_{Uh,\pi}) = \frac{1}{n_h}\left(1-\frac{n_h}{N_h}\right)S^2_{ys_h} \tag{5.2.15}$$

respectivamente. Nótese que $\hat{\bar{y}}_{Uh,\pi}$ es insesgado para la media del estrato $\bar{y}_{Uh}$ de la característica de interés y, y que $\widehat{Var}_{MAS}(\hat{\bar{y}}_{Uh,\pi})$ es insesgado para $Var_{MAS}(\hat{\bar{y}}_{Uh,\pi})$.

Por el contrario del razonamiento que se tuvo en la estimación del total poblacional, **es equivocado pensar de la siguiente manera:**

> Si un estimador insesgado del total poblacional t_y es la suma de cada una de las estimaciones en los H estratos, entonces un estimador del promedio poblacional $\bar{y}_U$ será un promedio de los promedios estimados en los H estratos.

El anterior razonamiento es intuitivo pero es errado la siguiente razón:

$$\bar{y}_U \neq \frac{\bar{y}_{U_1} + \bar{y}_{U_2} + \ldots + \bar{y}_{U_H}}{H}$$

Es fácil verlo con nuestra población de ejemplo U en donde el primer estrato U_1 tiene una media igual a $\bar{y}_{U_1} = 67.5$, el segundo estrato U_2 tiene una media igual a $\bar{y}_{U_2} = 33.67$. Por tanto $(\bar{y}_{U_1} + \bar{y}_{U_2})/2 = 50.58$ mientras que la verdadera media poblacional es $\bar{y}_U = 47.2$.

Resultado 5.2.5. *Bajo un diseño de muestreo aleatorio simple sin reemplazo en el estrato h, un estimador insesgado de la media $\bar{y}_U$, su varianza y su varianza estimada están dados por*

$$\hat{\bar{y}}_{U,\pi} = \frac{1}{N}\hat{t}_{y,\pi} = \frac{1}{N}\sum_{h=1}^{H} N_h\hat{\bar{y}}_{Uh,\pi} \tag{5.2.16}$$

$$Var_{MAE}(\hat{\bar{y}}_{U,\pi}) = \frac{Var_{MAE}(\hat{t}_{y,\pi})}{N^2} = \frac{1}{N^2}\sum_{h=1}^{H}\frac{N_h}{n_h}\left(1-\frac{n_h}{N_h}\right)S^2_{yU_h} \tag{5.2.17}$$

$$\widehat{Var}_{MAE}(\hat{\bar{y}}_{U,\pi}) = \frac{\widehat{Var}_{MAE}(\hat{t}_{y,\pi})}{N^2} = \frac{1}{N^2}\sum_{h=1}^{H}\frac{N_h}{n_h}\left(1-\frac{n_h}{N_h}\right)S^2_{ys_h} \tag{5.2.18}$$

respectivamente. Nótese que $\hat{\bar{y}}_{U,\pi}$ es insesgado para la media poblacional $\bar{y}_{Uh}$ de la característica de interés y, y que $\widehat{Var}_{MAS}(\hat{\bar{y}}_{U,\pi})$ es insesgado para $Var_{MAE}(\hat{\bar{y}}_{U,\pi})$.

Intervalos de confianza

Al respecto Lohr (2000) afirma que un intervalo de $100(1-\alpha)\,\%$ de confianza para la media de una población está dado por

$$\hat{\bar{y}}_{U,\pi} \pm Z_{1-\frac{\alpha}{2}}\sqrt{Var_{MAE}(\hat{\bar{y}}_{U,\pi})} \tag{5.2.19}$$

si se cumple algunas de las siguientes condiciones

- El tamaño de muestra n_h en cada estrato h es grande.
- Existe una gran número de estratos.

Si las anteriores condiciones no pueden ser satisfechas, se prefiere utilizar el percentil de una distribución t-student con $N-H$ grados de libertad. Así, un intervalo de confianza para la media poblacional está dado por

$$\hat{\bar{y}}_{U,\pi} \pm t_{1-\frac{\alpha}{2},N-H}\sqrt{Var_{MAE}(\hat{\bar{y}}_{U,\pi})} \tag{5.2.20}$$

5.2.4 Asignación del tamaño de muestra

Tal vez, la parte más importante en el diseño de una encuesta es la determinación del tamaño de muestra. En muestreo estratificado, bajo la restricción de que el tamaño de la muestra general es n y de la existencia de H estratos fijos, se quiere determinar los tamaños de muestra n_h para cada estrato h de tal manera que se garantice la ganancia de precisión del estimador. Lehtonen & Pahkinen (2003) señalan que en investigaciones por muestreo reales, las cuales incluyen varias características de interés, es imposible lograr que la asignación de la muestra arroje ganancias en la eficiencia de manera global (para cada una de las características de interés).

Asignación proporcional

Se decide utilizar este tipo de asignación cuando la muestra debe ser representativa de la población de acuerdo al comportamiento de la información auxiliar. Lohr (2000) lo expresa de la siguiente manera

> *Al utilizar la asignación proporcional, la muestra se puede ver como una versión miniatura de la población.*

Si se define la **fracción de muestreo** como $f_h = n_h/N_h$ en el estrato h, entonces al utilizar la asignación proporcional la fracción de muestreo será la misma para todos los estratos, tal que $f_h = f$. Nótese que la probabilidad de inclusión de cualquier elemento en la población $\pi_k = f_h = f$ es constante y fija. De esta manera, cada unidad en la muestra representará el mismo número de elementos en la población, independientemente del estrato al que pertenezca.

Definición 5.2.2. *Un diseño de muestreo aleatorio estratificado tiene asignación proporcional si*

$$\frac{n_h}{N_h} = \frac{n}{N} \qquad h = 1, \ldots, H \tag{5.2.21}$$

Resultado 5.2.6. *Para un diseño de muestreo aleatorio estratificado con asignación proporcional, el estimador de Horvitz-Thompson del total poblacional t_y, su varianza y su varianza estimada están dados por:*

$$\hat{t}_{y,\pi} = \frac{N}{n}\sum_{k\in S} y_k \tag{5.2.22}$$

$$Var_{MAE}(\hat{t}_{y,\pi}) = \frac{N^2}{n}\left(1-\frac{n}{N}\right)\sum_{h=1}^{H}\frac{n_h}{n}S^2_{yU_h} \tag{5.2.23}$$

$$\widehat{Var}_{MAE}(\hat{t}_{y,\pi}) = \frac{N^2}{n}\left(1-\frac{n}{N}\right)\sum_{h=1}^{H}\frac{n_h}{n}S^2_{ys_h} \tag{5.2.24}$$

Prueba. Observando la relación de la definición anterior se tiene que

$$\begin{aligned}\hat{t}_{y,\pi} &= \sum_{h=1}^{H}\frac{N_h}{n_h}\sum_{k\in S_h} y_k\\ &= \frac{N}{n}\sum_{h=1}^{H}\sum_{k\in S_h} y_k\\ &= \frac{N}{n}\sum_{k\in S} y_k\end{aligned}$$

Para las varianzas se tiene que

$$\begin{aligned}\sum_{h=1}^{H}\frac{N_h^2}{n_h}\left(1-\frac{n_h}{N_h}\right)S^2_{yU_h} &= \sum_{h=1}^{H}\frac{N_h^2}{n_h^2}\left(1-\frac{n_h}{N_h}\right)n_h S^2_{yU_h}\\ &= \frac{N^2}{n^2}\left(1-\frac{n}{N}\right)\sum_{h=1}^{H} n_h S^2_{yU_h} = \frac{N^2}{n}\left(1-\frac{n}{N}\right)\sum_{h=1}^{H}\frac{n_h}{n}S^2_{yU_h}\end{aligned}$$

■

Asignación de Neyman

Jerzy Neyman en su artículo de 1934, discutía el problema de la selección de una muestra mediante métodos probabilísticos versus la selección de una muestra a

conveniencia. En ese artículo, él observa las grandes bondades de los dos métodos. Sin embargo, mostró que separando la población en subgrupos poblacionales que llamó estratos y tomando muestras aleatorias simples sin reemplazo, los límites del intervalo de confianza podían ser minimizados para un tamaño de muestra fijo. Este artículo fue fundamental en el uso del muestreo estratificado alrededor del mundo.

Neyman trató con el problema de minimizar la varianza $Var_{MAE}(\hat{t}_{y,\pi})$ del estimador de Horvitz-Thompson fijando el tamaño de muestra general n. Como lo mencionan Groves, Fowler, Couper, Lepkowski, Singer & R. (2004), bajo este método se producen las menores varianzas para la media muestral comparado con otras técnicas de asignación de tamaño de muestra. Para realizar esta asignación es necesario conocer los tamaños de muestra en cada estrato n_h tal que $\sum_{h=1}^{H} n_h = n$.

Resultado 5.2.7. *Bajo la asignación de Neyman, el tamaño de muestra que minimiza (5.2.11) está dado por*

$$n_h = n\frac{N_h S_{yU_h}}{\sum_{h=1}^{H} N_h S_{yU_h}} \tag{5.2.25}$$

donde $S_{yU_h} = \sqrt{S^2_{yU_h}}$

Prueba. La cantidad a minimizar es

$$\sum_{h=1}^{H}\frac{N_h^2}{n_h}\left(1-\frac{n_h}{N_h}\right)S^2_{yU_h}$$

sujeta a

$$\sum_{h=1}^{H} n_h = n$$

La ecuación de Lagrange se escribe como

$$\mathcal{L}(n_1,\ldots,n_h,\lambda) = \sum_{h=1}^{H}\frac{N_h^2}{n_h}\left(1-\frac{n_h}{N_h}\right)S^2_{yU_h} - \lambda\left(n-\sum_{h=1}^{H} n_h\right) \tag{5.2.26}$$

al anular las derivadas parciales se tiene

$$\frac{\partial\mathcal{L}}{\partial\lambda} = n - \sum_{h=1}^{H} n_h = 0 \tag{5.2.27}$$

$$\frac{\partial\mathcal{L}}{\partial n_h} = -\frac{N_h^2}{n_h^2}S^2_{yU_h} + \lambda = 0 \tag{5.2.28}$$

De (5.2.28) se tiene que

$$n_h = \frac{N_h}{\sqrt{\lambda}}S_{yU_h} \tag{5.2.29}$$

Reemplazando en (5.2.27)

$$\sum_{h=1}^{H} n_h = n = \frac{\sum_{h=1}^{H} N_h S_{yU_h}}{\sqrt{\lambda}}$$

Por tanto,

$$\sqrt{\lambda} = \frac{1}{n}\sum_{h=1}^{H} N_h S_{yU_h} \tag{5.2.30}$$

Por último, reemplazando en (5.2.29) se tiene que

$$n_h = n\frac{N_h S_{yU_h}}{\sum_{h=1}^{H} N_h S_{yU_h}}$$

Es posible mostrar que la matriz de segundas derivadas parciales es definida positiva para los valores que satisfacen las restricciones. Así se concluye que lo valores de n_h dados por este resultado minimizan la varianza del estimador de Horvitz-Thompson bajo un tamaño de muestra fijo. ■

Por supuesto, es necesario conocer las varianzas de la característica de interés en cada estrato para poder utilizar este método. Con respecto a la asignación de Neyman se tienen problemas de redondeo, en este caso es recomendable redondear al entero más próximo. Sin embargo, la expresión (5.2.25) puede llevar a la situación en donde $n_h > N_h$. En este caso, se realiza un censo en el estrato en donde la anterior relación se presente y luego se restablece el cálculo de n_h para los demás estratos. Cuando se realiza un censo en un estrato, debido a la asignación de Neyman, o al diseño logístico de la encuesta, ese estrato es llamado **estrato de inclusión forzosa**.

Aunque utilizar este método puede guiar a ganancias en la eficiencia de la estrategia de muestreo, Groves, Fowler, Couper, Lepkowski, Singer & R. (2004) señalan las siguientes debilidades de la asignación de Neyman:

- Al estimar proporciones no se tienen buenos resultados. Dado a que se requiere que las proporciones tengan grandes diferencia entre los estratos. En la vida práctica esta situación no se tiene en la mayoría de ocasiones.
- Por construcción, este método funciona bien bajo el supuesto de que sólo existe una característica de interés. Cuando se tiene trabaja en encuesta multi-propósito no se tiene una reducción de varianza para todas las características de interés incluidas en la investigación.

Asignación óptima

Este es un método más general que la asignación de Neyman. Si al interior de algún estrato, existe una gran variabilidad, el anterior método de asignación induce un mayor tamaño de muestra en el estrato. Como lo expresa Lohr (2000) en el sector empresarial, por ejemplo, las ventas de las compañías grandes tienen un mucho mayor dispersión que las ventas de las micro-empresas.

Sin embargo si, como en la mayoría de situaciones prácticas, se cuenta con recursos económicos limitados para la realización del estudio. Y dado un capital, se quiere minimizar la varianza de la estrategia de muestreo, se debe realizar otro tipo de asignación. Por lo tanto definiendo la siguiente función de costos

$$C = \sum_{h=1}^{H} n_h C_h \tag{5.2.31}$$

En donde C_h es el costo de obtener la información para las características de interés de un elemento seleccionado y perteneciente al estrato h y C es el costo total de la realización del estudio. Luego, si se quiere distribuir la selección de elemento entre los estratos dado un costo fijo C, de manera que se minimice la varianza del estimador de Horvitz-Thompson, se debe utilizar la asignación óptima.

Resultado 5.2.8. *Bajo la asignación óptima, el tamaño de muestra que minimiza la función de coste está dado por*

$$n_h = \frac{C}{\sqrt{c_h}} \frac{N_h S_{yU_h}}{\sum_{i=1}^{H} N_i \sqrt{c_i} S_{yU_i}} \tag{5.2.32}$$

Prueba. Resulta inmediata al utilizar un razonamiento similar a la demostración del resultado de la asignación de Neyman. Es posible mostrar que la matriz de segundas derivadas parciales es definida positiva para los valores que satisfacen las restricciones. Así se concluye que lo valores de n_h dados por este resultado minimizan la varianza del estimador de Horvitz-Thompson bajo un coste fijo. ■

La expresión de la asignación óptima lleva a las siguientes conclusiones. En un determinado estrato, se debe seleccionar una muestra de tamaño grande sí:

- El tamaño del estrato N_h es grande y la recolección de la información en el estrato es más barata.
- El estrato tiene una gran dispersión con respecto a la característica de estudio. En este caso, se extrae una muestra más grande para compensar la heterogeneidad dentro del estrato.

5.2.5 Estimación en dominios

La estimación por dominios se caracteriza por el desconocimiento de la pertenencia de las unidades poblacionales al dominio. Es decir, para conocer cuáles unidades de la población pertenecen al dominio, es necesario realizar el proceso de medición. Sin embargo, existe un símil entre los estratos y los dominios y es que los dos dividen la población en subgrupos poblacionales. Por un lado, mientras que el conocimiento a priori de la pertenencia de los elementos poblacionales a los estratos ayuda a mejorar la eficiencia de la estimación en la etapa de diseño de la encuesta. Por otro lado, el precio que se debe pagar por el desconocimiento de la pertenencia de los elementos poblacionales a los dominios resulta alto.

Uno de los propósitos del diseño de muestreo estratificado es reducir la varianza de las estimaciones para la característica de interés. Esto se cumple en el caso en donde el comportamiento de la característica de interés (como se verá en las próximas secciones) toma valores promedio distintos en cada estrato. Sin embargo, en la estimación de proporciones para dominios no se garantiza que la anterior regal se cumpla.

Ahora, al multiplicar la variable de pertenencia al dominio z_{dk} dada por (3.2.22) por el valor de la característica de interés y_k, se crea una nueva variable y_{dk} dada por $y_{dk} = z_{dk} y_k$, y una vez construida se utilizan los principios del estimador de Horvitz-Thompson para hallar un estimador insesgado del total de la característica

de interés en el dominio U_d, el tamaño absoluto del dominio y la media de la característica en el dominio. Por supuesto, antes de obtener las estimaciones a nivel poblacional, es necesario aunque no suficiente, obtener las estimaciones de los dominios en los estratos.

Estimación del total en un dominio

Resultado 5.2.9. *Bajo muestreo aleatorio estratificado, el estimador de Horvitz-Thompson para el total del dominio t_{yhd} en el estrato h, su varianza y su varianza estimada están dados por*

$$\hat{t}_{yhd,\pi} = \frac{N_h}{n_h}\sum_{S_h} y_{hdk} \tag{5.2.33}$$

$$Var(\hat{t}_{yhd,\pi}) = \frac{N_h^2}{n_h}\left(1-\frac{n_h}{N_h}\right)S^2_{y_{dU_h}} \tag{5.2.34}$$

$$\widehat{Var}(\hat{t}_{yhd,\pi}) = \frac{N_h^2}{n_h}\left(1-\frac{n_h}{N_h}\right)S^2_{y_{ds_h}} \tag{5.2.35}$$

respectivamente. y_{hdk} es el valor de la nueva característica y_{dk} en el h-ésimo estrato. $S^2_{y_{dU_h}}$ y $S^2_{y_{ds_h}}$ denotan el estimador de la varianza de los valores de la característica de interés y_{dk} en el estrato U_h y en la muestra s_h seleccionada de dicho estrato, respectivamente.

Resultado 5.2.10. *Bajo muestreo aleatorio estratificado, el estimador de Horvitz-Thompson para el total del dominio t_{yd} en la población, su varianza y su varianza estimada están dados por*

$$\hat{t}_{yd,\pi} = \sum_{h=1}^{H}\frac{N_h}{n_h}\sum_{S_h} y_{hdk} \tag{5.2.36}$$

$$Var(\hat{t}_{yd,\pi}) = \sum_{h=1}^{H}\frac{N_h^2}{n_h}\left(1-\frac{n_h}{N_h}\right)S^2_{y_{dU_h}} \tag{5.2.37}$$

$$\widehat{Var}(\hat{t}_{yd,\pi}) = \sum_{h=1}^{H}\frac{N_h^2}{n_h}\left(1-\frac{n_h}{N_h}\right)S^2_{y_{ds_h}} \tag{5.2.38}$$

Nótese que en la expresión $S^2_{y_{dU_h}}$ los valores que intervienen son: los de la característica de interés, si el elemento pertenece al dominio, y ceros si el elemento no pertenece al dominio, lo mismo sucede con $S^2_{y_{ds_h}}$. Por tanto, las anteriores expresiones de varianza van a tomar valores grandes por la inclusión de los ceros; éste es el precio que se debe pagar por el desconocimiento de la pertenencia de los elementos a los dominios.

Estimación de la media de un dominio

Resultado 5.2.11. *Bajo muestreo aleatorio estratificado, el estimador de Horvitz-Thompson para la media de la característica de interés en un dominio $\bar{y}_{dU_h}$ en el*

estrato h, su varianza y su varianza estimada están dados por

$$\hat{\bar{y}}_{dU_h,\pi} = \frac{\hat{t}_{yhd,\pi}}{N_{hd}} \tag{5.2.39}$$

$$Var(\hat{\bar{y}}_{dU_h,\pi}) = \frac{1}{N_{hd}^2} Var(\hat{t}_{yhd,\pi}) \tag{5.2.40}$$

$$\widehat{Var}(\hat{\bar{y}}_{dU_h,\pi}) = \frac{1}{N_{hd}^2} \widehat{Var}(\hat{t}_{yhd,\pi}) \tag{5.2.41}$$

Resultado 5.2.12. *Bajo muestreo aleatorio estratificado, el estimador de Horvitz-Thompson para la media de la característica de interés en un dominio $\bar{y}_d$ en la población, su varianza y su varianza estimada están dados por*

$$\hat{\bar{y}}_{d,\pi} = \frac{\hat{t}_{yd,\pi}}{N_d} \tag{5.2.42}$$

$$Var(\hat{\bar{y}}_{d,\pi}) = \frac{1}{N_d^2} Var(\hat{t}_{yd,\pi}) \tag{5.2.43}$$

$$\widehat{Var}(\hat{\bar{y}}_{d,\pi}) = \frac{1}{N_d^2} \widehat{Var}(\hat{t}_{yd,\pi}) \tag{5.2.44}$$

Para poder utilizar los anteriores resultados, es necesario conocer de antemano el valor del tamaño absoluto del dominio en cada estrato N_{hd} y el valor del tamaño absoluto del dominio en la población N_d.

Estimación del tamaño absoluto de un dominio

Resultado 5.2.13. *Bajo muestreo aleatorio estratificado, el estimador de Horvitz-Thompson para el tamaño absoluto de un dominio N_{hd} en el estrato h, su varianza y su varianza estimada están dados por*

$$\hat{N}_{hd,\pi} = \frac{N_h}{n_h} \sum_{S_h} z_{dk} \tag{5.2.45}$$

$$Var(\hat{N}_{hd,\pi}) = \frac{N_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right) S^2_{z_{dU_h}} \tag{5.2.46}$$

$$\widehat{Var}(\hat{N}_{hd,\pi}) = \frac{N_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right) S^2_{z_{ds_h}} \tag{5.2.47}$$

respectivamente, con $S^2_{z_{dU_h}}$ y $S^2_{z_{ds_h}}$ el estimador de la varianza de los valores de la característica de interés z_{dk} en el estrato U_h y en la muestra s_h seleccionada de dicho estrato.

Resultado 5.2.14. *Bajo muestreo aleatorio estratificado, el estimador de Horvitz-Thompson para el tamaño absoluto de un dominio N_d en la población, su varianza y su varianza estimada están dados por*

$$\hat{N}_{d,\pi} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{S_h} z_{dk} \tag{5.2.48}$$

$$Var(\hat{N}_{d,\pi}) = \sum_{h=1}^{H} \frac{N_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right) S^2_{z_{dU_h}} \tag{5.2.49}$$

$$\widehat{Var}(\hat{N}_{d,\pi}) = \sum_{h=1}^{H} \frac{N_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right) S^2_{z_{ds_h}} \tag{5.2.50}$$

respectivamente.

Nótese que en la expresión $S^2_{z_{dU_h}}$ los valores que intervienen son unos, si el elemento pertenece al dominio U_d, y ceros si el elemento no pertenece al dominio, lo mismo sucede con $S^2_{y_d s}$.

Estimación del tamaño relativo de un dominio

Resultado 5.2.15. *Bajo muestreo aleatorio estratificado, el estimador de Horvitz-Thompson para el tamaño relativo de un dominio P_{hd} en el estrato h, su varianza y su varianza estimada están dados por*

$$\hat{P}_{hd,\pi} = \frac{1}{N_h} \hat{N}_{hd,\pi} = \frac{1}{n_h} \sum_{S_h} z_{dk} = \frac{n_{hd}}{n_h} \tag{5.2.51}$$

$$Var(\hat{P}_{hd,\pi}) = \frac{1}{n_h} \left(1 - \frac{n_h}{N_h}\right) S^2_{z_{dU_h}} \tag{5.2.52}$$

$$\widehat{Var}(\hat{P}_{hd,\pi}) = \frac{1}{n_h} \left(1 - \frac{n_h}{N_h}\right) S^2_{z_{ds_h}} \tag{5.2.53}$$

Resultado 5.2.16. *Bajo muestreo aleatorio estratificado, el estimador de Horvitz-Thompson para el tamaño relativo de un dominio P_d en la población, su varianza y su varianza estimada están dados por*

$$\hat{P}_{d,\pi} = \frac{\hat{N}_{d,\pi}}{N} = \frac{1}{N} \sum_{h=1} \frac{N_h}{n_h} \sum_{S_h} z_{dk} \tag{5.2.54}$$

$$Var(\hat{P}_{d,\pi}) = \frac{1}{N^2} \sum_{h=1}^{H} \frac{N_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right) S^2_{z_{dU_h}} \tag{5.2.55}$$

$$\widehat{Var}(\hat{P}_{d,\pi}) = \frac{1}{N^2} \sum_{h=1}^{H} \frac{N_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right) S^2_{z_{ds_h}} \tag{5.2.56}$$

5.2.6 El efecto de diseño

Lehtonen & Pahkinen (2003) plantean que la eficiencia del diseño de muestreo estratificado depende fuertemente de la proporción de variación total en cada estrato. Es decir, utilizando los resultados del análisis de varianza, tenemos el siguiente resultado:

Resultado 5.2.17. *Suponga que la población se divide en h grupos, de tal forma que existen N_h elementos por grupo y el tamaño poblacional toma la forma $N = \sum_{h=1}^{H}$, entonces*

$$(N-1)S^2_{y_U} = \underbrace{\sum_U (y_k - \bar{y}_U)^2}_{SCT} = \underbrace{\sum_{h=1}^{H}\sum_{U_h} (y_{hk} - \bar{y}_{U_h})^2}_{SCD} + \underbrace{\sum_{h=1}^{H} N_h (\bar{y}_{U_h} - \bar{y}_U)^2}_{SCE} \tag{5.2.57}$$

Empíricamente observando la construcción de la varianza del estimador de Horvitz-Thompson en la ecuación (5.2.11) se puede inferir que para tener una varianza pequeña, la variación al interior de los estratos debe ser pequeña. Es decir, los estratos deben ser homogéneos por dentro. Cada esquema de asignación de muestras arroja resultados diferentes en cuanto a la eficiencia se refiere. En esta sección se considera el esquema de asignación de muestra proporcional dado por la definición 5.2.2. en donde la varianza del estimador de Horvitz-Thompson está dada por la siguiente expresión:

$$Var_{MAE}(\hat{t}_{y,\pi}) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right)\sum_{h=1}^{H} W_h S^2_{yU_h} \tag{5.2.58}$$

donde $S^2_{yU_h}$ es la varianza de la característica de interés en el estrato h y $W_h = \frac{n_h}{n}\frac{N_h}{N}$. Con un poco de álgebra se llega al siguiente resultado.

Resultado 5.2.18. *Bajo un diseño de muestreo aleatorio simple sin reemplazo con asignación proporcional, la varianza del estimador de Horvitz-Thompson toma la siguiente forma*

$$Var_{MAS}(\hat{t}_{y,\pi}) \cong \frac{N^2}{n}\left(1 - \frac{n}{N}\right)\sum_{h=1}^{H} W_h \left[S^2_{yU_h} + (\bar{y}_{U_h} - \bar{y}_U)^2\right] \tag{5.2.59}$$

Prueba.

$$\begin{aligned}
(N-1)S^2_{yU_h} &= \sum_U (y_k - \bar{y}_U)^2 && (5.2.60)\\
&= \sum_{h=1}^{H}\sum_U (y_{hk} - \bar{y}_U)^2 && (5.2.61)\\
&= \sum_{h=1}^{H}\sum_{U_h} (y_{hk} - \bar{y}_{U_h})^2 + \sum_{h=1}^{H} N_h (\bar{y}_{U_h} - \bar{y}_U)^2 && (5.2.62)\\
&= \sum_{h=1}^{H} (N_h - 1)S^2_{yU_h} + \sum_{h=1}^{H} N_h (\bar{y}_{U_h} - \bar{y}_U)^2 && (5.2.63)
\end{aligned}$$

Por tanto

$$S^2_{yU_h} \cong \sum_{h=1}^{H} \frac{N_h}{N} \left[S^2_{yU_h} + (\bar{y}_{U_h} - \bar{y}_U)^2\right] \tag{5.2.64}$$

$$= \frac{N^2}{n}\left(1 - \frac{n}{N}\right) \sum_{h=1}^{H} W_h \left[S^2_{yU_h} + (\bar{y}_{U_h} - \bar{y}_U)^2\right] \tag{5.2.65}$$

■

Resultado 5.2.19. *El efecto de diseño en el muestreo aleatorio simple sin reemplazo con asignación proporcional está dado por*

$$Deff \cong \frac{\sum_{h=1}^{H} W_h S^2_{yU_h}}{\sum_{h=1}^{H} W_h \left[S^2_{yU_h} + (\bar{y}_{U_h} - \bar{y}_U)^2\right]} \tag{5.2.66}$$

$$\tag{5.2.67}$$

$$\cong \frac{\textit{Varianza dentro de los estratos}}{\textit{Varianza Total}} \tag{5.2.68}$$

Ahora, intuitivamente tenemos que

Varianza Total = Varianza dentro + Varianza entre

Por tanto se concluye que, casi siempre, esta estrategia de muestreo arrojará mejores resultados que una estrategia aleatoria simple.

5.2.7 Marco y Lucy

En investigaciones anteriores (que no ha utilizado información auxiliar), el gobierno ha establecido que la característica SPAM no es un motor de desarrollo, en cuanto a ingreso neto se refiere, en las empresas del sector industrial. Lo anterior puede obedecer a razones de tipo gerencial o a la cultura organizacional de las empresas en el sector. Por supuesto, el *modus operandi* del gerente de marca y las estrategias de posicionamiento de marca en el mercado varían de acuerdo a la productividad y tamaño de la empresa. De hecho, no es posible, por cuestiones financieras y logísticas, que una empresa de muy baja productividad utilice los medios publicitarios que una empresa de alto nivel pueda utilizar. Las empresas de alto nivel han dispuesto una parte de sus ganancias en la reinversión publicitaria en medios masivos de comunicación. Las empresas de bajo nivel no pueden hacer esto porque sus márgenes de ganancia no se prestan para pautar en esta clase de medios.

Por lo anterior, cada estrategia de mercadeo es diferente, entre otras, porque cada cliente de cada empresa es diferente de acuerdo al nivel de productividad en el sector industrial. Es decir, los clientes de las empresas grandes son clientes que se caracterizan porque realizan pedidos de varios millones de dólares, y los clientes de las empresas pequeñas se caracterizan por ser empresas emergentes y, en algunos casos, personas naturales independientes, por tanto el margen de ganancias en cada nivel del sector empresarial es muy distinto.

Sin embargo, independientemente del tipo de cliente e incluso del nivel de la empresa en el sector industrial, existe una herramienta que todas las empresas en el sector industrial pueden utilizar: el envío de publicidad directa mediante el uso del correo electrónico. Por supuesto, en países no desarrollados, en las empresas pequeñas, una vez más ya sea por el tipo de gerencia o cultura organizacional o incluso por cuestiones financieras, no existe la infraestructura ni la capacitación para establecer este tipo de publicidad no convencional.

Bajo estos antecedentes, el gobierno está dispuesto a brindar planes de financiamiento a todas las empresas del sector industrial, por lo que ha planeado una nueva investigación acerca de los hábitos y usos del SPAM en las empresas del sector industrial para observar el desarrollo que el sector ha tenido gracias a este medio. La figura 5.1. muestra el comportamiento de las tres características de interés para el gobierno. Se nota que existe una mayor variabilidad en las empresas que pertenecen al nivel **Grande**, mientras que la variabilidad en los niveles **Mediano** y **Pequeño** es menor. Más aún, el comportamiento promedio de las variables de interés es distinto en cada estrato. Esto implica que utilizar un diseño de muestreo aleatorio estratificado sería una buena decisión si se quiere ganar en precisión.

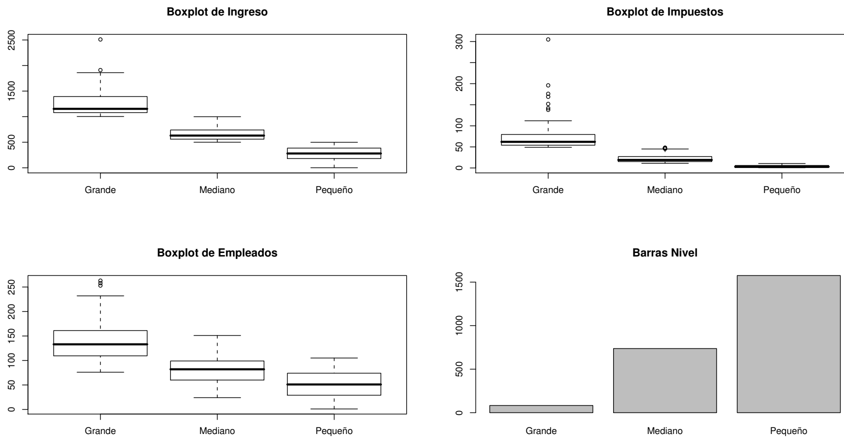


Figura 5.1: *Boxplot de las características de interés en cada nivel industrial.*

Por supuesto, el gobierno ha creado un plan de políticas con la promesa de beneficiar al electorado. Si el gobierno corrobora la hipótesis, por medio del presente estudio, de la influencia del SPAM en el crecimiento del algún nivel del sector industrial, entonces buscará planes de capacitación y financiamiento para que las empresas de los niveles **Mediano** y **Pequeño** crezcan, se estabilicen y fomenten la creación de nuevos empleos y el tributo a las entidades gubernamentales pertinentes y, que las empresas del nivel **Grande** no desciendan de nivel sino que se expandan no sólo nacionalmente sino que también en el ámbito internacional a donde también puede llegar la publicidad SPAM en cuestión de micro segundos.

Para esta nueva investigación, el gobierno ha proveído un marco de muestreo que además de contener la ubicación y la identificación de todas las empresas de todos lo niveles industriales, también adjunta el tipo de empresa, a saber: **Grande,**

Media, Pequeña. El tipo de empresa será tomada como variable de estratificación para el diseño del plan muestral.

Estimación del tamaño de muestra

El gobierno está decidido en implementar un plan de capacitación a las empresas del sector industrial y ha pedido que el diseño de muestreo sea representativo de la población en cuanto a la característica de estratificación: Nivel. Está razón nos lleva a utilizar la asignación proporcional para la estimación del tamaño de muestra. Se planea la realización de $n = 400$ encuestas, por tanto $f = \frac{400}{2396} = 0,167$. Esto implica la realización de $n_1 = 14$ encuestas de empresas grandes, $n_2 = 123$ encuestas en empresas medianas y $n_3 = 263$ encuestas en empresas pequeñas.

Para la selección de la muestra, se debe cargar el marco de muestreo en el ambiente de R. Con la variable de estratificación Nivel se determinan los tamaños de cada uno de las estratos que se debe convertir en un vector de tamaño $H = 3$, así `N <- c(N1,N2,N3)`, lo mismo se debe hacer con los tamaños de muestra en cada estrato, se deben convertir en vector así `n <- c(n1,n2,n3)`.

```
> data(Lucy)
> attach(Lucy)
> N1 <- summary(Level)[[1]]
> N2 <- summary(Level)[[2]]
> N3 <- summary(Level)[[3]]
> N1;N2;N3
[1] 83
[1] 737
[1] 1576
> N <- c(N1,N2,N3)
> n1 <- 14
> n2 <- 123
> n3 <- 263
> n <- c(n1,n2,n3)
```

Utilizando la función `S.STSI` del paquete `TeachingSampling` es posible seleccionar una muestra aleatoria simple en cada uno de los tres estratos. Esta función consta de tres argumentos. El primero: `Estrato`, es la variable de estratificación que indica la pertenencia de todos y cada uno de los $\sum_{h=1}^{H} N_h = N$ individuos de la población. El segundo argumento: `N`, un vector de tamaño H que indica los tamaños de cada estrato en la población. El último argumento: `n`, un vector de tamaño H que indica los tamaños de muestra en cada estrato. El resultado de la función es un conjunto de índices que, aplicados a la población, permite la obtención de la muestra estratificada.

```
> sam <- S.STSI(Level,N,n)
> muestra <- Lucy[sam,]
> attach(muestra)
> muestra
         ID Ubication  Level Zone
2336 AB1126    c25k59    Big    A
```

```
2321 AB1039    c25k44    Big    A
...
1676 AB1095    c17k92 Medium    A
2088  AB724     c23k9 Medium    D
...
322  AB1514     c4k25  Small    B
1576  AB962    c16k91  Small    E
```

La muestra realizada (seleccionada) es de tamaño 400 y está dividida en cada uno de los tres estratos. Una vez que la selección de los elementos es efectuada, se necesita obtener la información mediante una encuesta a cada una de las empresas del sector industrial. Nótese que en este punto, la realización de un muestreo estratificado tiene ventajas logísticas. Lo anterior es evidente cuando se decide que el cuestionario será enviado vía correo electrónico a cada una de las 14 empresas del nivel **Grande**. Por tanto, la realización de esta entrevista arroja ventajas financieras enormes pues el envío de un correo electrónico no supone mayor gasto. Para la realización de la encuesta en el nivel **Mediano** se ha decidido contratar a una agencia de correos postales y, de esa forma, hacer llegar mediante correo certificado un cuestionario con la respectiva encuesta. No se aplica el mismo medio logístico que en las empresas grandes pues se sabe que no todas las empresas medianas tienen una dirección de correo electrónico actualizada, lo que no sucede en el estrato grande. Para obtener la información del sector industrial se ha decidido enviar encuestadores entrenados para el trabajo. Lo anterior se hace dado que los propietarios de las empresas pequeñas son reacios a responder las cartas certificadas y mucho menos responden el correo electrónico dado que tienen compromisos operativos que atender.

Una vez conseguida la información de cada una de las 400 empresas seleccionadas, se procede a estimar las cantidades de interés. Para esto se utiliza la función `E.STSI` del paquete `TeachingSampling`. Esta función consta de cuatro parámetros muestrales, a saber: `Estrato`, es la variable de estratificación que indica la pertenencia de todos y cada uno de los $\sum_{h=1}^{H} n_h = n$ individuos seleccionados en la muestra, `N` y `n`, los vectores del tamaño de la población y muestra estratificada respectivamente y `estima` conteniendo el valor de la(s) característica(s) de interés en cada uno de los elementos seleccionados.

```
> estima <- data.frame(Income, Employees, Taxes)
> E.STSI(Level,N,n,estima)

, , Income
                 Grande      Mediano      Pequeño        Total
Total est  1.024279e+05 4.866417e+05 4.389550e+05 1.028025e+06
Varianza   1.677372e+07 5.331612e+07 1.190085e+08 1.890984e+08
cve        3.998493e+00 1.500445e+00 2.485244e+00 1.337644e+00

, , Employees
                 Grande      Mediano      Pequeño        Total
Total est  1.176229e+04 6.035011e+04 8.367182e+04 1.557842e+05
Varianza   5.130655e+05 3.150449e+06 5.470307e+06 9.133822e+06
cve        6.089683e+00 2.941089e+00 2.795289e+00 1.940005e+00
```

```
, , Taxes
                  Grande      Mediano      Pequeño        Total
Total est  5.887071e+03 1.617206e+04  5968.425856 2.802755e+04
Varianza   2.068935e+05 2.788990e+05 79853.628890 5.656461e+05
cve        7.726345e+00 3.265564e+00     4.734646 2.683411e+00
```

La función `E.STSI` arroja la estimación de cada una de las características de interés discriminada por cada estrato y el gran total así como también la varianza estimada y el coeficiente de variación estimado. Nótese que en cuestión de ingreso, se estima que el estrato grande produce un 10 %, el estrato mediano un 47 % y el estrato pequeño un 43 % del ingreso neto del sector industrial. Un resultado similar se observa con las restantes características de interés. Nótese que los coeficientes de variación estimados en cada estrato son, en algunos casos elevados[4]; sin embargo, el coeficiente de variación para el total es bajo.

En la siguiente tabla se muestran los resultados particulares para este ejercicio. Se puede notar que la estratificación arroja buenos resultados con coeficientes de variación menores a los que arrojaría una muestra aleatoria simple. Esto se debe a que las variables de interés presentan, en promedio, un comportamiento diferente en cada estrato.

Tabla 5.1: *Muestreo aleatorio estratificado: estimación de los totales de las características de interés.*

Variable	Total poblacional	Total estimado	cve %	Desv. %
Ingreso	1035217	1028025	1.33	-0.69
Empleados	151950	155784	1.94	2.52
Impuestos	28654	28027	2.68	-2.19

La función `Domains` contenida en el paquete `TeachingSampling` se utiliza para obtener las variables indicadoras z_{dk} para cada dominio, el único argumento de la función es un vector de pertenencia de cada individuo. En este caso, el vector de pertenencia es SPAM, la salida de esta función es una matriz de unos y ceros, en donde cada columna está dicotomizada. Existen tantas columnas como subgrupos poblacionales, y en cada columna el número uno implica la pertenencia del elemento al dominio y cero la no pertenencia del elemento al dominio.

```
> Dominios <- Domains(SPAM)

> SPAM.si <- Dominios[,2]*estima
> SPAM.no <- Dominios[,1]*estima
```

Para estimar el tamaño absoluto de cada dominio, lo único que se debe hacer es multiplicar la matriz de características de interés (en este caso, la matriz llamada `estima`) por cada columna de la matriz resultante de la dicotomización. Utilizando la función `E.STSI` en la matriz resultante de la dicotomización obtenemos las estimación de los tamaños absolutos de cada dominio. En este caso, se estima que

[4] El coeficiente de variación es más alto a medida que las estimaciones estén más discriminadas en grupos.

1390 empresas ya están utilizando otras técnicas de publicidad como el SPAM, mientras que las restantes 1006 no lo están haciendo. Además en cada uno de los tres estratos existen más empresas que están utilizando el SPAM que las que no lo están haciendo y es interesante que en el estrato de las empresas pequeñas por cada 2 empresas que no utilizan el SPAM existen 3 que sí lo hacen.

Nótese que la varianza de cada estimación sigue siendo la misma, puesto que los valores de esta característica de interés son ceros y uno y, por tanto, la estructura de varianza resulta idéntica en cada caso.

```
> E.STSI(Level,N,n,Dominios)

, , no
              Grande     Mediano      Pequeño        Total
Total est   35.57143  323.560976   647.178707  1006.311111
Varianza   107.88697  913.500219  1911.452916  2932.840105
cve         29.20002    9.341104     6.755501     5.381606

, , si
              Grande     Mediano      Pequeño        Total
Total est   47.42857  413.439024   928.821293  1389.688889
Varianza   107.88697  913.500219  1911.452916  2932.840105
cve         21.90002    7.310429     4.707059     3.896966
```

Esta claro que existe una tendencia en el sector industrial de publicidad virtual mediante el envío de SPAM por correo electrónico.

Las siguientes cifras son las verdaderamente importantes pues muestran que las empresas en cada uno de los tres estratos que utilizan SPAM tienen mayores ingresos, emplean a más gente y contribuyen con una mayor cantidad de dinero en cuanto a impuestos se refiere, esto se da porque hay más empresas que utilizan el SPAM de las que no lo hacen. Se debe tener en cuenta que al interior de los subgrupos (estratos y dominios) el coeficiente de variación es alto en parte por la discriminación y en parte porque la varianza de las nuevas variables.

```
> E.STSI(Level,N,n,SPAM.si)

, , Income
                 Grande      Mediano      Pequeño        Total
Total est  5.687871e+04 2.749070e+05 2.495593e+05 5.813450e+05
Varianza   1.632250e+08 4.342108e+08 2.124806e+08 8.099164e+08
cve        2.246175e+01 7.579918e+00 5.840982e+00 4.895377e+00

, , Employees
                 Grande      Mediano      Pequeño        Total
Total est  6.663714e+03 3.420759e+04 4.621335e+04 8.708465e+04
Varianza   2.510026e+06 7.939778e+06 8.154733e+06 1.860454e+07
cve        2.377512e+01 8.237241e+00 6.179273e+00 4.952994e+00
```

```
, , Taxes
               Grande       Mediano       Pequeño         Total
Total est   3195.500  9.209504e+03   3349.749049  1.575475e+04
Varianza  586655.629  6.132710e+05  72204.334593  1.272131e+06
cve           23.969  8.503348e+00      8.021754  7.159032e+00

> E.STSI(Level,N,n,SPAM.no)
, , Income
                  Grande      Mediano      Pequeño        Total
Total est   4.554921e+04 2.117347e+05 1.893957e+05 4.466796e+05
Varianza    1.848997e+08 4.140719e+08 2.071224e+08 8.060941e+08
cve         2.985295e+01 9.610497e+00 7.598775e+00 6.356189e+00

, , Employees
                  Grande      Mediano      Pequeño        Total
Total est   5.098571e+03 2.614253e+04 3.745846e+04 6.869956e+04
Varianza    2.348367e+06 7.424198e+06 8.324746e+06 1.809731e+07
cve         3.005623e+01 1.042263e+01 7.702568e+00 6.192315e+00

, , Taxes
               Grande       Mediano       Pequeño         Total
Total est    2691.57    6962.55285   2618.676806  1.227280e+04
Varianza   720262.40  541369.25806  63436.086720  1.325068e+06
cve            31.53      10.56765      9.618034  9.379406e+00
```

Las siguientes tablas resumen las estimaciones de los dominios utilizando el anterior enfoque. Nótese que el valor de los coeficientes de variación es alto puesto que se trata de estimación en subgrupos poblacionales en donde el tamaño de muestra es aleatorio.

Tabla 5.2: *Estimación en el dominio SPAM SI.*

Variable	Total poblacional	Total estimado	cve %	Desv. %
Nd	1495	1390	3.89	-7,02
Ingreso	637265	581345	4.89	-8.77
Empleados	92350	87084	4.95	-5.70
Impuestos	18061	15754	7.15	-12.77

Tabla 5.3: *Estimación en el dominio SPAM NO.*

Variable	Total poblacional	Total estimado	cve %	Desv. %
Nd	937	1006	7.36	5.38
Ingreso	397752	446679	12.24	6.35
Empleados	59600	68699	15.27	6.19
Impuestos	10593	12272	15.85	9.37

Esto muestra que la utilización del SPAM puede ser una estrategia de crecimiento en el sector industrial. Ahora, pensando un poco en la eficiencia de la estrategia de muestreo, consideremos la siguiente tabla de análisis de varianza para calcular el efecto de diseño usando el resultado 5.2.19.

```
> data(Lucy)
> attach(Lucy)


> anova(lm(Income~Level))

Analysis of Variance Table

Response: Income
            Df    Sum Sq   Mean Sq F value    Pr(>F)
Level        2 129744890  64872445  3789.5 < 2.2e-16 ***
Residuals 2393  40965657     17119
```

El efecto de diseño estaría dado por $\frac{17119}{71278} = 0.24$. Por ello la eficiencia de la estrategia es cuatro veces mayor que una estrategia simple. Es interesante que un diseño tan sencillo como el simple en cada estrato con un tamaño de muestre pequeño arroje estos buenos resultados.

Nótese que como N_d es desconocido, para obtener otro tipo de estimación (aunque no la varianza ni el c.v.e) de la media de la característica de interés en cada dominio, podemos utilizar un estimador alternativo dado por

$$\widehat{y}_{S_d} = \frac{\hat{t}_{yd,\pi}}{\hat{N}_{d,\pi}} = \frac{\sum_S y_{dk}}{z_{dk}} = \frac{\sum_{S_d} y_k}{n_d}$$

Para ello, simplemente tomamos las estimaciones t_{yd} y las dividimos por la estimación de N_d.

Otro tipo de asignación

Suponga que el gobierno quiera hacer una encuesta con las características y magnitudes de la anterior, pero con un limitante importante: el dinero, el gobierno tiene un presupuesto de 7000 dólares para la realización del estudio. Además de esto, el gobierno quiere que el método usado para la recolección de la información sea clásico. Es decir, un encuestador debe ir a cada empresa y realizar el cuestionario. Este caso es muy frecuente en encuestas de mercadeo, en donde se quiere lograr buenas estimaciones pero no se dispone de muchos recursos financieros ni logísticos.

En este caso se ha averiguado que las varianzas de la variable ingreso son las siguientes 64398, 16081, 15142 en los estratos **Grande, Mediano y Pequeño** respectivamente. Además realizar una sola encuesta en el estrato de las empresas grandes cuesta alrededor de 40 dólares, una encuesta en el estrato de las empresas medianas cuesta 20 dólares y una entrevista en el estrato de las empresas pequeñas cuesta 15 dólares. Nótese la diferencia de precios en cada estrato, esto se debe a que es necesaria la contratación de encuestadores de alto perfil para las entrevistas en el estrato de las empresas grandes.

Utilizando la asignación óptima y el resultado 5.2.8. se tienen los tamaños de muestra en cada estrato, dados por la tabla anterior, que minimizan la varianza del estimador de Horvitz-Thompson con la restricción del costo total del estudio, 7000 dólares. Nótese que $\sum_{h=1}^{3} n_h C_h = 7000$.

Tabla 5.4: *Estimación del tamaño de muestra.*

Estrato	Coste	Nh	S2yuh	nh
Grande	40	83	64398	18
Mediano	20	737	16081	112
Pequeño	15	1576	15142	269

5.3 Diseño de muestreo estratificado PPT

Como se vio en la sección anterior, la ganancia de precisión al utilizar un diseño de muestreo estratificado es importante. Sin embargo, los resultados pueden mejorarse al utilizar una característica continua auxiliar x_k bien relacionada con la característica de interés y_k en cada estrato. Así, es posible estimar el parámetro de interés mediante el estimador de Hansen-Hurwitz con una varianza pequeña. De hecho, entre mejor correlación exista entre y y x, asumiendo que el comportamiento promedio de la variable de interés es diferente en cada estrato, menor varianza tendrá el estimador de Hansen-Hurwitz.

En este caso, el marco de muestreo debe tener dos características auxiliares: una variable de estratificación y la información auxiliar continua, ambas disponibles para cada elemento en todos los estratos. Se supone que el diseño de muestreo dentro de cada estrato es con reemplazo y, de esta manera, se selecciona una muestra de tamaño m_h en cada estrato h $(h = 1, \ldots, H)$. Cada elemento de $k \in U_h$ tiene probabilidad de selección igual a

$$p_k = \frac{x_k}{t_{xh}} \quad \text{si } k \in U_h \tag{5.3.1}$$

con t_{xh} el total poblacional de la característica auxiliar x en el estrato U_h. Es importante verificar que en cada estrato se cumpla

$$\sum_{U_h} p_k = 1 \quad \text{para cada } h = 1, \ldots, H, \tag{5.3.2}$$

por tanto

$$\sum_{h=1}^{H} \sum_{U_h} p_k = H \tag{5.3.3}$$

Ahora, en cada estrato U_h de tamaño N_h se selecciona una muestra s_h con reemplazo de tamaño m_h, por tanto la cardinalidad del soporte en el estrato U_h está dada por

$$\#Q_h = \binom{N_h + m_h - 1}{m_h} \tag{5.3.4}$$

El soporte general estratificado, se define como la unión de los soportes en cada uno de los estratos U_h.

$$Q^H = \left\{ \bigcup_{h=1}^{H} s_h \quad | \quad s_h \in Q_h \right\}. \tag{5.3.5}$$

5.3.1 Algoritmos de selección

En la selección de las muestras PPT con reemplazo en cada estrato es posible utilizar los algoritmos de muestreo dados en el capítulo 3, de tal forma que los siguientes pasos se deben realizar:

- Separar la población en H estratos mediante la variable de estratificación.
- En cada estrato U_h, seleccionar una muestra PPT con reemplazo. Los algoritmos utilizados en la selección de la muestra dentro de cada estrato pueden ser los métodos acumulativo total o el método de Lahiri.
- Cada una de las H selecciones es realizada de manera independiente.

5.3.2 El estimador de Hansen-Hurwitz

Con los anteriores condicionamiento, se utiliza el estimador de Hansen-Hurwitz para estimar de manera insesgada al parámetro de interés t_y con ayuda de información auxiliar continua en cada estrato U_h.

Resultado 5.3.1. *Si los elementos dentro del estrato U_h son seleccionados con reemplazo, de acuerdo a probabilidades de selección tales que $\sum_{U_h} p_k = 1$, basados en x_k, el valor de una característica auxiliar continua, entonces el estimador de Hansen-Hurwitz del total poblacional t_{yh}, su varianza y su varianza estimada están dados por:*

$$\hat{t}_{yh,p} = \frac{t_{xh}}{m_h} \sum_{\substack{i=1 \\ k \in S_h}}^{m_h} \frac{y_{ki}}{x_{ki}} \tag{5.3.6}$$

$$Var_{PPT}(\hat{t}_{yh,p}) = \frac{1}{m_h} \sum_{U_h} p_k \left(\frac{y_k}{p_k} - t_{yh} \right)^2 \tag{5.3.7}$$

$$\widehat{Var}_{PPT}(\hat{t}_{yh,p}) = \frac{1}{m_h(m_h-1)} \sum_{\substack{i=1 \\ k \in S_h}}^{m_h} \left(\frac{y_{ki}}{p_{ki}} - \hat{t}_{yh,p} \right)^2 \tag{5.3.8}$$

respectivamente, con p_k dados por (5.3.1). Nótese que $\hat{t}_{yh,p}$ es insesgado para el total t_{yh} de la característica de interés y, y que $\widehat{Var}_{PPT}(\hat{t}_{yh,p})$ es insesgado para $Var_{PPT}(\hat{t}_{yh,p})$.

Resultado 5.3.2. *Para un diseño de muestreo estratificado con selección de unidades PPT en cada estrato, el estimador de Hansen-Hurwitz del total poblacional t_y, su varianza y su varianza estimada están dados por:*

$$\hat{t}_{yh,p} = \sum_{h=1}^{H} \frac{t_{xh}}{m_h} \sum_{\substack{i=1 \\ k \in S_h}}^{m_h} \frac{y_{ki}}{x_{ki}} \tag{5.3.9}$$

$$Var_{EPPT}(\hat{t}_{yh,p}) = \sum_{h=1}^{H} \frac{1}{m_h} \sum_{U_h} p_k \left(\frac{y_k}{p_k} - t_{yh} \right)^2 \tag{5.3.10}$$

$$\widehat{Var}_{EPPT}(\hat{t}_{yh,p}) = \sum_{h=1}^{H} \frac{1}{m_h(m_h-1)} \sum_{\substack{i=1\\k\in S_h}}^{m_h} \left(\frac{y_{ki}}{p_{ki}} - \hat{t}_{yh,p}\right)^2 \qquad (5.3.11)$$

respectivamente. Nótese que $\hat{t}_{y,p}$ es insesgado para el total t_y de la característica de interés y, y que $\widehat{Var}_{EPPT}(\hat{t}_{y,p})$ es insesgado para $Var_{EPPT}(\hat{t}_{y,p})$.

Ejemplo 5.3.1. Para nuestra población de ejemplo U particionada en 2 estratos como en el capítulo anterior, existen por un lado $\binom{N_1+m_1-1}{m_1} = 6$ posibles muestras con reemplazo de tamaño $m_1 = 2$ en el primer estrato y por el otro lado $\binom{N_2+m_2-1}{m_2} = 2$ posibles muestras con reemplazo de tamaño $m_2 = 1$ en el segundo estrato. Utilizando la característica auxiliar x, realice el cálculo léxico-gráfico del estimador de Hansen-Hurwitz y compruebe el insesgamiento y la varianza.

5.3.3 Marco y Lucy

En la pasada sección, supusimos que el marco de muestreo contenía, además de la ubicación e identificación de todas las empresas del sector industrial, una variable de estratificación llamada Nivel que agrupa a las empresas de acuerdo a su capacidad de producción industrial. Es lógico pensar que el comportamiento promedio de las características de interés es diferente en cada estrato. Siendo así los resultados obtenidos son más precisos que al realizar un plan de muestreo simple, además de obtener las estimaciones de las características de interés anidadas en los estratos.

En esta ocasión, la construcción del marco de muestreo ha logrado incluir además de la variable de estratificación Nivel una información auxiliar continua, particularmente se supone que se tiene conocimiento del valor de ingreso declarado en el último año fiscal para cada empresa del sector industrial.

Con este generoso marco de muestreo es claro que las estimaciones serán más precisas. Aunque vale la pena preguntarse si la eficiencia de las estimaciones mejorará notablemente con estas dos variables auxiliares. Se utilizará la asignación proporcional, como en la sección pasada, para hacer los resultados comparables. No olvide que en cada estrato la selección de las muestras se hace con reemplazo.

```
> data(Lucy)
> attach(Lucy)
> m1 <- 14
> m2 <- 123
> m3 <- 263
> m <- c(m1,m2,m3)
```

La función `S.STPPS(E,x,m)` se utiliza para la extracción de las H muestras con reemplazo en cada estrato. Los argumentos de la función son los siguientes: `E`, la variable de estratificación en la población entera, en este caso particular es Nivel. `x`, un vector de información auxiliar continua conteniendo cada uno de los valores en la población, en este caso particular es `Income`. `m`, un vector conteniendo H tamaños de muestra para cada estrato.

La función `S.STPPS(E,x,m)` divide el marco de muestreo en H estratos y en cada uno de ellos selecciona una muestra con reemplazo de acuerdo a probabilidades

de selección dadas por (5.3.1)[5]. El resultado de la función es en dos vías: por una parte, la función devuelve los índices de los elementos seleccionados con reemplazo en cada estrato y, por otra, devuelve el vector de probabilidades de selección de los elementos en la muestra. Cada una de las anteriores salidas es de tamaño $m = \sum_{h=1}^{H} m_h$. Para este ejercicio el resultado de la función se ha guardado en el objeto `res`, la muestra en el objeto `sam` y el vector de probabilidades de selección en la muestra se ha guardado en el objeto `pk`.

```
> res <- S.STPPS(Level,Income,m)
> sam <- res[,1]
> pk <- res[,2]
> muestra <- Lucy[sam,]
> attach(muestra)
> muestra
            ID Ubication  Level Zone Income Employees Taxes SPAM
2344   AB1199    c25k67     Big    B   1150        88  62.0  yes
2369.1  AB748    c25k92     Big    D   1300       172  68.0  yes
2320   AB1038    c25k43     Big    A   1044        90  52.0  yes
1957    AB575    c20k76 Medium    C    550        35  14.0  yes
...
```

Aplicando los índices obtenido en `sam` al marco de muestreo, obtenemos la información para realizar el proceso de recolección de datos. Cuando la información es recolectada se creará un archivo de datos conteniendo cada uno de los valores de la(s) característica(s) de interés en la muestra seleccionada. Esta archivo es adjuntado a R mediante la función `attach`.

La etapa de estimación se realiza con la función `E.STPPS(y,pk,m,E)` del paquete `TeachingSampling` cuyos argumentos son cuatro y cada uno de ellos contiene información a nivel de la muestra y nada más que de la muestra: `y`, el archivo de datos conteniendo cada uno de los valores de la(s) característica(s) de interés en la muestra seleccionada, en este caso particular será el `data frame estima`. `pk` el vector de probabilidades de selección resultante de aplicar la función `S.STPPS` en la etapa de selección de muestra, en esta caso particular guardado como `pk <- res[,2]`. `m`, un vector conteniendo H tamaños de muestra para cada estrato, en este caso dado por `m <- c(m1,m2,m3)`. `E`, la variable de estratificación en la muestra, en este caso particular es `Level` en la muestra no en la población.

La función `E.STPPS` arroja la estimación de cada una de las características de interés discriminada por cada estrato y el gran total así como también la varianza estimada y el coeficiente de variación estimado. También arroja las estimaciones de los tamaños de los estratos $\hat{N}_h$ y del tamaño de la población total dado por $\hat{N} = \sum_{h=1}^{H} \hat{N}_h$.

```
> estima <- data.frame(Income, Employees, Taxes)
> E.STPPS(estima, pk, m, Level)
```

[5]Esta función trata cada estrato como una población separada de modo que la suma de las probabilidades de selección en cada estrato suman uno y en toda la población suman H.

```
, , 1
               Grande      Mediano      Pequeño        Total
Total est  85.5308339 739.22157976 1.517982e+03 2.342734e+03
Varianza    0.2224841   0.22805284 9.655518e-01 1.416089e+00
cve         0.5514762   0.06460158 6.473233e-02 5.079515e-02

, , Income
                Grande      Mediano      Pequeño        Total
Total est  1.03706e+05 4.873510e+05 4.441600e+05 1.035217e+06
Varianza   6.77756e-26 8.033922e-27 2.170687e-27 7.798029e-26
cve        2.51034e-16 1.839170e-17 1.048960e-17 2.697497e-17

, , Employees
                Grande      Mediano      Pequeño        Total
Total est  1.17441e+04 6.084383e+04 7.898202e+04 1.515700e+05
Varianza   3.70116e+03 3.547822e+03 5.067016e+03 1.231601e+04
cve        5.18021e-01 9.789587e-02 9.012554e-02 7.321868e-02


, , Taxes
                 Grande      Mediano      Pequeño        Total
Total est  6040.0336101 1.630756e+04 6.234842e+03 2.858244e+04
Varianza   3145.7857046 1.461127e+02 1.518736e+01 3.307086e+03
cve           0.9285926 7.412334e-02 6.250514e-02 2.011979e-01
```

Nótese que las estimaciones dentro de los estratos tienen un coeficiente de variación muy pequeño al igual que la estimación para la población total. La siguiente tabla muestra los resultados para este ejercicio particular.

Tabla 5.5: *Muestreo estratificado PPT: estimación de los totales de las características de interés.*

Variable	Total poblacional	Total estimado	cve %	Desv. %
Ingreso	1035217	1035217	0.00	0.00
Empleados	151950	151570	0.07	-0.25
Impuestos	28654	28582	0.20	-0.25

Es notable la ganancia en eficiencia de esta estrategia de muestreo, no hay mucho que decir al respecto. Simplemente se deben agotar hasta los últimos recursos para poder estratificar la población y aplicar un diseño de muestreo PPT en cada estrato, siempre y cuando la característica de interés esté bien correlacionada en cada estrato con la información auxiliar.

5.4 Ejercicios

5.1 Demuestre teóricamente o refute mediante un contraejemplo las siguientes afirmaciones:

(a) Para aplicar un diseño de muestreo estratificado se pide que los estratos no se traslapen. La anterior condición es necesaria para estimar la varianza del estimador.
(b) La necesidad de estratificar surge siempre de razones administrativas.
(c) Siempre un diseño de muestreo estratificado es de menor varianza que un diseño de muestreo que no incluya estratos.
(d) En un diseño de muestreo estratificado, la estimación del promedio poblacional es el promedio de las estimaciones de los totales en cada estrato.
(e) Explique una ventaja técnica para estratificar
(f) Explique una ventaja logística para estratificar
(g) Exponga detalladamente un ejemplo en donde para diferentes estratos se proponen diferentes diseños de muestreo.

5.2 Escriba las fórmulas del estimador del total y del estimador de la varianza para los siguientes diseños de muestreo. Defina cada término y notación que utilice en las fórmulas.

(a) Diseño estratificado con tres estratos: uno de inclusión forzosa, otro con diseño PPT y otro con diseño MAS.
(b) Diseño estratificado con dos estratos: uno de inclusión forzosa, otro con diseño sistemático.
(c) Diseño estratificado con cuatro estratos: uno de inclusión forzosa, otro con diseño bernoulli, con diseño MAS con reemplazo y otro con diseño Poisson proporcional a una característica de inormación auxiliar.
(d) Diseño estratificado con tres estratos: todos con diseño πPT

5.3 Realice el ejercicio lexicográfico del Ejemplo 5.1.1.

5.4 Realice el ejercicio lexicográfico del Ejemplo 5.2.2.

5.5 Realice el ejercicio lexicográfico del Ejemplo 5.3.1.

5.6 Suponga una población de cuatro elementos $U = \{1, 2, 3, 4\}$ cuyos valores para la caraterística de interés son $y_1 = y_2 = 0$, $y_3 = 1$, $y_4 = -1$. En primer lugar, calcule la varianza del estimador de la media poblacional para un diseño de muestreo aleatorio simple con tamaño de muestra $n = 2$. Luego, calcule la varianza del estimador de la media poblacional para un diseño de muestreo con dos estratos $U_1 = \{1, 2\}$ y $U_2 = \{3, 4\}$ si dentro de cada estrato se planea un diseño aleatorio simple de tamaño uno. ¿Cuál varianza resultó ser más grande?. Explique

5.7 Suponga que una población de municipios se divide en dos estratos, uno urbano y el otro rural. De todas los municipios en la población, siete ($N_1 = 7$) son ciudades y los restantes veiticinco ($N_2 = 25$) son distritos rurales. Se decide que se usará un diseño de muestreo estratificado de tamaño total $n = 8$. Teniendo en cuenta la siguiente tabla, determine tamaños de muestra en cada estrato de acuerdo a la afijación proporcional, afijación de Neyman y afijación óptima.

	Estrato rural	Estrato urbano	Población Total
Media	283	1146	472
Desv. Est.	331	1318	743
Tamaño	25	7	32
Costo por encuesta	5 pesos	2 pesos	3 pesos

5.8 Calcule el estimador del total poblacional, el estimador de la media poblacional, sus respectivos c.v.e. e intervalos de confianza para una estrategia de muestreo que utiliza el estimador de Horvitz - Thompson y un diseño de muestreo aleatorio estratificado ($H = 2$). El tamaño del primer estrato es de $N_1 = 105$ y el del segundo estrato es de $N_2 = 19$. Para el estrato uno, se seleccionó una muestra de $n_1 = 11$ elementos y para el estrato dos, se seleccionó una muestra de $n_2 = 4$ elementos. Use la siguiente información:

Estrato h	$\sum_{s_h} y_k$	$\sum_{s_h} y_k^2$
1	1099	21855
2	3446	1822736

Capítulo 6

Muestreo de conglomerados

> En encuestas complejas, los grupos poblacionales de elementos que se forman naturalmente como barrios, municipios o escuelas pueden ser tratados como unidades de muestreo. Este tipo de esquemas de muestreo ayudan a aumentar el tamaño de muestra manteniendo el costo de la encuesta.
>
> Risto Lehtonen (2004)

Las estrategias de muestreo para elementos tienen un común denominador: el marco de muestreo y su prolija identificación y ubicación de los elementos poblacionales, de todos y cada uno de ellos. Cabe resaltar que en la práctica no es muy común el uso de diseños de muestreo que seleccionen muestras de elementos directamente. Lo anterior se debe más a cuestiones financieras y logísticas que a problemas de eficiencia estadística. Piense en lo siguiente: cada investigación requiere un marco de muestreo. Son miles de millares las investigaciones realizadas al año y deberían existir tantos marcos de muestreo como investigaciones realizadas. Por cuestiones de tipo logístico la consecución de un marco de muestreo de elementos es muy costosa porque implicaría realizar un censo, enumerando, identificando y ubicando a cada elemento de la población y esto es, por supuesto, algo utópico.

Pensando en el más sencillo de los diseños de muestreo, el costo financiero de realizar un estudio mediante un diseño de muestreo aleatorio simple es muy elevado. Por ejemplo suponga que se desea realizar un estudio para evaluar la calidad de vida de las personas en un determinado país. Si llegara a existir un marco de muestreo de elementos, realizar (o seleccionar) una muestra aleatoria simple demandaría la contratación de un encuestador por cada persona encuestada, puesto que la dispersión geográfica natural de los elementos seleccionados en la muestra aleatoria simple sería demasiado alta.

En el caso anterior, aunque se tuviera un marco de muestreo de elementos, el costo financiero de realizar una muestra aleatoria sería demasiado alto. Una forma de realizar muestras probabilísticas a falta de un marco de muestreo de elementos es seleccionar conglomerados[1] de elementos y realizar el proceso de medición en cada conglomerado. Cochran (1977) plantea que, por cuestiones logísticas, es más eficiente seleccionar una muestra de 20 bloques de hogares, cada bloque con 30

[1] Agrupación natural de objetos

hogares, que seleccionar una muestra aleatoria de 600 hogares. En el primer caso sólo se necesitaría la presencia de un encuestador por bloque, mientras que en el segundo, posiblemente, se necesite la presencia de muchos más encuestadores.

Siempre que se desee seleccionar una muestra probabilística se debe tener un marco de muestreo de manera obligatoria, en los casos en donde se carece de marco muestral es necesario construir uno. Sin embargo, el costo financiero y logístico de levantar un marco de muestreo para elementos es muy alto, en la mayoría de ocasiones. Una forma de construir marcos de muestreo de bajo costo es mediante la aplicación de un diseño de muestreo por conglomerados. Estos conglomerados tienen la ventaja de ser agrupaciones de elementos que se forman de manera natural y además existen entidades gubernamentales que se ocupan de registrar y actualizar la lista de conglomerados existentes en cada sector. Por ejemplo, existe una entidad encargada de la actualización de los sectores cartográficos de una ciudad, existe una entidad encargada de la actualización de los negocios en un sector, existe una entidad que recopila la información concerniente a la ubicación de las escuelas, etc. Para cada entidad existe también un registro de estas aglomeraciones y este será el marco de muestreo que se utilizará en la etapa de diseño.

Por lo tanto, el marco de muestreo contendrá la ubicación e identificación de cada uno de los conglomerados de elementos existentes en la población. Con este marco de conglomerados, se aplica un diseño de muestreo y una muestra es seleccionada. Cada conglomerado seleccionado en la muestra es visitado y el proceso de medición se realiza para todos los elementos pertenecientes al mismo. Entonces, si el conglomerado seleccionado es una sección cartográfica de la ciudad, se aplicará la encuesta a todos y cada uno de los elementos que conforman la sección. Si el conglomerado seleccionado es una escuela, se aplicará el instrumento de medición a todos y cada uno de los alumnos de la escuela. En otras palabras, se realiza un censo en cada conglomerado que haya sido seleccionado en la muestra.

Por supuesto, existe una ganancia significativa en términos operativos, logísticos y financieros. Sin embargo, esta ganancia tiene un precio... el precio a pagar está dado en términos de eficiencia estadística de la estrategia de muestreo. Revisando un poco el proceso de aglomeración, hay que tener en cuenta que los conglomerados de elementos tienden, en la mayoría de los casos, a ser homogéneos con respecto a los valores de la característica de interés y. Lo anterior se da porque la agrupación se realiza de forma natural, es decir lo hogares, las secciones cartográficas, las villas, las escuelas, las prisiones, etc. tienden a formarse de manera natural y homogénea.

De esta manera, la pérdida de eficiencia estadística es causada por el efecto de conglomerado que conlleva la selección de unidades homogéneas que no contienen información nueva sino, de alguna manera, repetida. ¿Qué nueva información se obtiene, acerca de la población, al añadir un nuevo elemento del mismo conglomerado en la muestra?

Entre más grande sea el tamaño de la sub-muestra en los conglomerados, entonces más grande será el efecto de diseño. Si dentro de cada conglomerado, el comportamiento de la característica de interés y reflejará el comportamiento estructural de la misma en la población, entonces la eficiencia de una estrategia de muestreo por conglomerados sería similar a la de una muestra aleatoria simple. Pero, en la práctica, la homogeneidad interna de los conglomerados aumenta el error de muestreo.

Un error, por desgracia demasiado frecuente, entre los investigadores neófitos es analizar una muestra por conglomerados como una muestra aleatoria simple[2]. En general se tienen lo siguientes comentarios acerca del muestreo por conglomerados:

- Utilizamos muestreo por conglomerados sí:
  1. La construcción de un marco de muestreo de elementos es muy difícil, muy costosa o imposible de conseguir. Enumerar abejas, enumerar clientes, enlistar árboles en un sector, enlistar hogares en los barrios conglomerados (dispersión geográfica, reducción de costos).
  2. La población objetivo se encuentra muy dispersa (geográficamente) o aparece en agrupaciones naturales: familias, escuelas, etc.
- Los elementos individuales de una población sólo participan en la muestra si pertenecen a un conglomerado incluido en la muestra.
- El muestreo estratificado aumenta la precisión de las estimaciones, mientras que el muestreo por conglomerados tiende a disminuirla. Es un precio que se paga al no poseer un marco de muestreo definido para los elementos de la población objetivo.
- Al obtener una muestra de elementos que pertenecen a un conglomerado repetimos la información del conglomerado (dada la agrupación natural). Lo ideal es conseguir información nueva en cada individuo, por lo anterior se pierde precisión en las estimaciones.

6.1 Fundamentos teóricos y notación

Suponga que la población de elementos

$$U = \{1, ..., k, ..., N\}.$$

se divide en N_I sub-grupos poblacionales, llamados **conglomerados** y denotados como $U_I = \{U_1, \ldots, U_{N_I}\}$.

La población de conglomerados estará dada, sin pérdida de generalidad, por

$$U_I = \{1, \ldots, N_I\}.$$

Estos definen una partición de la población en tal forma que

1. $U = \bigcup_{i=1}^{N_I} U_i$
2. $U_i \bigcap U_j = \emptyset$ para todo $i \neq j$

El número de unidades N_i en el conglomerado i-ésimo se llama **tamaño del conglomerado** tal que

$$N = \sum_{i=1}^{N_I} N_i,$$

[2]No es prudente, ni correcto analizar una muestra por conglomerados como si fuera una muestra aleatoria simple porque los errores estándar serán mayores y la interpretación de los resultados será errónea.

donde N es el tamaño de la población U. Con la población dividida en N_I conglomerados, los parámetros poblacionales de interés pueden escribirse como:

1. El total poblacional,

$$t_y = \sum_{k \in U} y_k = \sum_{i=1}^{N_I} \sum_{k \in U_i} y_k = \sum_{i=1}^{N_I} t_{yi} \tag{6.1.1}$$

   donde $t_{yi} = \sum_{k \in U_i} y_k$ es el total del i-ésimo conglomerado.

2. La media poblacional,

$$\bar{y}_U = \frac{\sum_{k \in U} y_k}{N} = \frac{1}{N} \sum_{i=1}^{N_I} \sum_{k \in U_i} y_k = \frac{1}{N} \sum_{i=1}^{N_I} N_i \bar{y}_i \tag{6.1.2}$$

   donde $\bar{y}_i = \dfrac{1}{N_i} \sum_{k \in U_i} y_k$ es la media del i-ésimo conglomerado.

El esquema general del diseño de muestreo por conglomerados está definido de la siguiente forma

1. Seleccionar una muestra probabilística s_I[3] de conglomerados de la población U_I mediante un diseño de muestreo tal que

$$Pr(S_I = s_I) = p_I(s_I) \qquad \text{para todo } s_I \in Q_I. \tag{6.1.3}$$

   donde Q_I es el soporte conteniendo todas las posibles muestras de conglomerados.

2. Todos y cada uno de los elementos pertenecientes a los conglomerados seleccionados son observados y medidos.

El tamaño de la muestra aleatoria de conglomerados está dado por

1. $n(S_I) = n_I$ si la muestra es de tamaño fijo, $n(S_I)$ si la muestra es de tamaño variable

2. $n(S_I) = m_I$ si la muestra es seleccionada con reemplazo

La muestra aleatoria de elementos viene caracterizada por

$$S = \bigcup_{i \in S_I} U_i \tag{6.1.4}$$

y el tamaño de la muestra[4] de elementos por

[3]Nótese que si s_I representa la muestra realizada de conglomerados, entonces S_I representa la muestra aleatoria la cual es una variable aleatoria.

[4]Dado que, por lo general, el tamaño de los conglomerados varía, se tiene que $n(S)$ es generalmente aleatorio incluso si $n(S_I)$ es de tamaño fijo.

$$n(S) = \sum_{i \in S_I} N_i \tag{6.1.5}$$

Si es posible construir o definir un soporte Q_I, también será posible definir (al menos teóricamente) un soporte general Q de elementos conteniendo las posibles muestras de elementos pertenecientes a los conglomerados seleccionados.

Ejemplo 6.1.1. Nuestra población ejemplo U dada por

$$U = \{\textbf{Yves, Ken, Erik, Sharon, Leslie}\}$$

se divide en tres conglomerados de la siguiente forma

$$U_1 = \{\textbf{Yves, Ken}\}$$

el segundo conformado por

$$U_2 = \{\textbf{Erik,Sharon}\}$$

y el último conglomerado dado por

$$U_3 = \{\textbf{Leslie}\}$$

Es claro que, en este caso particular, se tienen $N_I = 3$ conglomerados de tamaño diferentes. De esta manera, la población de conglomerados queda definida por

$$U_I = \{U_1, U_2, U_3\}$$

Suponga que se selecciona una muestra s_I de conglomerados de tamaño $n_I = 2$. La definición del soporte Q_I en R se hace mediante el uso de la función `Support` del paquete `TeachingSampling` aplicada a la información a nivel de los conglomerados de la siguiente manera.

```
> U   <-  c("Yves", "Ken", "Erik", "Sharon", "Leslie")
> U1  <-  c("Yves", "Ken")
> U2  <-  c("Erik", "Sharon")
> U3  <-  c("Leslie")
> UI  <-  c("U1","U2","U3")

> N1  <-  length(U1)
> N2  <-  length(U2)
> N3  <-  length(U3)

> ty1  <-  sum(32,34)
> ty2  <-  sum(46,89)
> ty3  <-  sum(35)

> tyI  <-  c(ty1,ty2,ty3)
```

```
> ty   <-  sum(ty1,ty2,ty3)

> NI=3
> nI=2

> QI  <-  Support(NI,nI,UI)
> QI

$m
     [,1] [,2]
[1,]    1    2
[2,]    1    3
[3,]    2    3

$muestras
     [,1] [,2]
[1,] "U1" "U2"
[2,] "U1" "U3"
[3,] "U2" "U3"
```

Suponga que se selecciona una muestra s_I de conglomerados de tamaño $m_I = 2$. Defina el soporte Q_I de conglomerados y el soporte Q de elementos si la selección de la muestra es con reemplazo.

6.1.1 El estimador de Horvitz-Thompson

Nótese que en el esquema general del muestreo por conglomerados, se utiliza un diseño de muestreo para la selección de los conglomerados en la muestra. Este diseño de muestreo $p_I(s_I)$ puede ser cualquiera de los diseños vistos en los capítulos anteriores, aplicados a la selección, esta vez no de elementos, sino de conglomerados. En general, dado el soporte Q_I, $p_I(s_I)$ puede ser:

- **Sin reemplazo:** si todas las posibles muestras en Q_I son sin reemplazo. Muestreo aleatorio simple, Bernoulli, Sistemático, Poisson, πPT o estratificado simple.
- **Con reemplazo:** si todas las posibles muestras en Q_I son con reemplazo. Muestreo aleatorio simple con reemplazo o muestreo PPT.
- **De tamaño fijo:** si todas las posibles muestras en Q tienen el mismo tamaño de muestra $n(S_I) = n_I$.

Nótese que el diseño de muestreo $p_I(s_I)$ induce probabilidades de inclusión sobre los conglomerados las cuales están definidas como sigue a continuación.

Definición 6.1.1. *La probabilidad de inclusión del conglomerado i-ésimo está dada por*

$$\pi_{Ii} = Pr(i \in S_I) = \sum_{s_I \ni i} p_I(s_I). \tag{6.1.6}$$

mientras que la probabilidad de inclusión de los conglomerados i-ésimo y j-ésimo están dadas por

$$\pi_{Iij} = Pr(i \in S_I \text{ y } j \in S_I) = \sum_{s_I \ni \ i \ y \ j} p_I(s_I). \tag{6.1.7}$$

respectivamente. Por supuesto, $\pi_{Iii} = \pi_{Ii}$.

Asimismo, debido a la naturaleza jerárquica de la agrupación de elementos en los conglomerados, el siguiente resultado muestra las probabilidades de inclusión al nivel de los elementos de la población.

Resultado 6.1.1. *La probabilidad de que el k-ésimo elemento, sea incluido en la muestra S está dada por*

$$\pi_k = \pi_{Ii} \quad \text{si } k \in U_i \tag{6.1.8}$$

Por otro lado, la probabilidad de inclusión de los elementos k-ésimo y l-ésimo está dada por

$$\pi_{kl} = \begin{cases} \pi_{Ii}, & \text{si } k,l \in U_i, \\ \\ \pi_{Iij}, & \text{si } k \in U_i, l \in U_j, i \neq j. \end{cases} \tag{6.1.9}$$

Una vez definidas las probabilidades de inclusión se define la estrategia de muestreo con el uso del estimador de Horvitz-Thompson, dado por el siguiente resultado

Resultado 6.1.2. *Bajo un diseño de muestreo por conglomerados, el estimador de Horvitz-Thompson para el total t_y, su varianza y su varianza estimada están dados por*

$$\hat{t}_{y,\pi} = \sum_{i \in S_I} \frac{t_{yi}}{\pi_{Ii}} \tag{6.1.10}$$

$$Var_1(\hat{t}_{y,\pi}) = \sum\sum_{U_I} \Delta_{Iij} \frac{t_{yi}}{\pi_{Ii}} \frac{t_{yj}}{\pi_{Ij}}. \tag{6.1.11}$$

$$\widehat{Var}_1(\hat{t}_{y,\pi}) = \sum\sum_{S_I} \frac{\Delta_{Iij}}{\pi_{Iij}} \frac{t_{yi}}{\pi_{Ii}} \frac{t_{yj}}{\pi_{Ij}} \tag{6.1.12}$$

respectivamente, con $\Delta_{Iij} = \pi_{Iij} - \pi_{Ii}\pi_{Ij}$ y t_{yi} el total del i-ésimo conglomerado seleccionado. Nótese que $\hat{t}_{y,\pi}$ es insesgado para t_y y que $\widehat{Var}_1(\hat{t}_{y,\pi})$ es insesgado para $Var_1(\hat{t}_{y,\pi})$.

Prueba. Para el estimador, se tiene que

$$\begin{aligned} \hat{t}_{y,\pi} &= \sum_{k \in S} \frac{y_k}{\pi_k} \\ &= \sum_{i \in S_I} \sum_{k \in U_i} \frac{y_k}{\pi_k} \\ &= \sum_{i \in S_I} \frac{1}{\pi_{Ii}} \sum_{k \in U_i} y_k \\ &= \sum_{i \in S_I} \frac{t_{yi}}{\pi_{Ii}}. \end{aligned}$$

Para el cálculo de la varianza es necesario verificar que

$$\Delta_{kl} = \Delta_{Iij} = \begin{cases} \pi_{Ii} - \pi_{Ii}^2, & \text{si } k,l \in U_i; \\ \pi_{Iij} - \pi_{Ii}\pi_{Ij}, & \text{si } k \in U_i,\ l \in U_j \text{ y } i \neq j \end{cases} \tag{6.1.13}$$

Entonces se tiene que

$$
\begin{aligned}
Var_1(\hat{t}_{y,\pi}) &= \sum_{k\in U}\sum_{l\in U}\Delta_{kl}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}\\
&= \sum_{i\in U_I}\sum_{k\in U_i}\sum_{j\in U_I}\sum_{l\in U_j}\Delta_{kl}\frac{y_k}{\pi_{Ii}}\frac{y_l}{\pi_{Ij}}\\
&= \sum_{i\in U_I}\sum_{j\in U_I}\Delta_{Iij}\frac{1}{\pi_{Ii}}\frac{1}{\pi_{Ij}}\sum_{k\in U_i}y_k\sum_{l\in U_j}y_l\\
&= \sum\sum_{U_I}\Delta_{Iij}\frac{t_{yi}}{\pi_{Ii}}\frac{t_{yj}}{\pi_{Ij}}
\end{aligned}
$$

Se procede análogamente para la estimación de la varianza. ■

Resultado 6.1.3. *Si el diseño de muestreo $p_I(s_I)$ es de tamaño fijo, la varianza del estimador de Horvitz-Thompson y su varianza estimada toman la siguiente forma*

$$Var_2(\hat{t}_{y,\pi}) = -\frac{1}{2}\sum\sum_{U_I}\Delta_{Iij}\left(\frac{t_{yi}}{\pi_{Ii}}-\frac{t_{yj}}{\pi_{Ij}}\right)^2 \tag{6.1.14}$$

$$\widehat{Var}_2(\hat{t}_{y,\pi}) = -\frac{1}{2}\sum\sum_{S_I}\frac{\Delta_{Iij}}{\pi_{Iij}}\left(\frac{t_{yi}}{\pi_{Ii}}-\frac{t_{yj}}{\pi_{Ijl}}\right)^2 \tag{6.1.15}$$

Nótese que $\widehat{Var}_2(\hat{t}_{y,\pi})$ es insesgado para $Var_2(\hat{t}_{y,\pi})$.

Prueba. La demostración de los anteriores resultados es inmediata siguiendo los lineamentos de la sección del estimador de Horvitz-Thompson del segundo capítulo y notando que $t_y = \sum_{U_I} t_{yi}$. ■

Al respecto de la construcción del estimador de Horvitz-Thompson bajo muestreo en conglomerados, Bautista (1998) deduce que

1. La eficiencia de la estrategia de muestreo toma su máximo valor cuando los valores $\dfrac{t_{yi}}{\pi_{Ii}}$ son constantes para todo $i = 1, \ldots, N_I$.
2. Cuando el diseño por conglomerados es tal que asigna probabilidades de inclusión idénticas a cada conglomerado, la estrategia pierde eficiencia, a menos que el comportamiento de los totales de cada conglomerado sea similar.

Los anteriores comentarios nos llevan a preferir diseños de muestreo que asignen probabilidades de inclusión proporcionales al tamaño del conglomerado. Para esto se debería disponer de información auxiliar continua disponible para toda la población U_I que estuviera bien correlacionada con los totales de la característica de interés en cada conglomerado t_{yi}. En otras palabras, nuestro marco de muestreo es de conglomerados; por tanto, si x representa la información auxiliar continua y t_{xi} el total de la información auxiliar en el i-ésimo conglomerado, la correlación entre t_{xi} y t_{yi} debería ser bastante fuerte y las probabilidades de inclusión de los conglomerados deberían corresponder a la siguiente forma funcional:

$$\pi_{Ii} = n_I \frac{t_{xi}}{t_x} \tag{6.1.16}$$

Ejemplo 6.1.2. Nuestra población ejemplo U_I dada por

$$U_I{=}\{U_1, U_2, U_3\}$$

Suponga que se selecciona una muestra s_I de conglomerados de tamaño $n_I = 2$ mediante un diseño de muestreo sin reemplazo tal que

$$p_I(s_I) = \begin{cases} 0.5, & \text{si } s_I{=}\{U_1, U_2\}, \\ 0.4, & \text{si } s_I{=}\{U_1, U_3\}, \\ 0.1, & \text{si } s_I{=}\{U_2, U_3\} \end{cases}$$

Mediante el siguiente ejercicio léxico-gráfico se comprueba el insesgamiento del estimador de Horvitz-Thompson en R. Para esto utilizamos las funciones `Ik` y `Pik` del paquete `TeachingSampling` a nivel de los conglomerados.

```
> p  <-  c(0.5, 0.4, 0.1)
> Ind <- Ik(NI,nI)
> data.frame(QI,p,Ind)

  X1 X2   p     I.1    I.2  I.3
1 U1 U2  0.5      1      1    0
2 U1 U3  0.4      1      0    1
3 U2 U3  0.1      0      1    1

> piI <- Pik(p,Ind)
> piI
[1] 0.9 0.6 0.5
```

De esta manera, la probabilidad de inclusión más alta la tiene el conglomerado U_1 y la más baja corresponde al conglomerado U_3. Con esto podemos calcular la estimación mediante el uso de la función `HT` del paquete `TeachingSampling`.

```
  X1 X2   p   pi.i pi.j  ty.i ty.j        est
1 U1 U2 0.5   0.9  0.6    66  135    298.3333
2 U1 U3 0.4   0.9  0.5    66   35    143.3333
3 U2 U3 0.1   0.6  0.5   135   35    295.0000

> sum(p*est)
[1] 236
```

Una vez más, nótese que en la estimación intervienen las cantidades de los conglomerados seleccionados en la muestra realizada.

Estimación de otros parámetros

Si el tamaño poblacional N es conocido, la media poblacional definida en (2.1.14) puede ser estimada con el estimador de Horvitz-Thompson.

Resultado 6.1.4. *En muestreo por conglomerado la media poblacional es estimada insesgadamente mediante el uso del estimador de Horvitz-Thompson*

$$\hat{\bar{Y}}_{\pi} = \frac{1}{N}\left(\hat{t}_{y,\pi}\right) \tag{6.1.17}$$

$$= \frac{1}{N}\sum_{i \in S_I}\frac{t_{yi}}{\pi_{Ii}} \tag{6.1.18}$$

Una de las razones por las que se utiliza el muestreo por conglomerados es la falta de un marco de muestreo para elementos. En este caso el desconocimiento del tamaño poblacional es muy típico. Sin embargo, utilizando los principios del estimador de Horvitz-Thompson, es posible estimar el tamaño de la población escribiéndolo como

$$N = \sum_{i \in U_I} N_i, \tag{6.1.19}$$

Luego, tenemos el siguiente resultado.

Resultado 6.1.5. *En muestreo por conglomerados el tamaño poblacional es estimado insesgadamente mediante el uso de la siguiente expresión*

$$\hat{N}_{\pi} = \sum_{i \in S_I}\frac{N_i}{\pi_{Ii}} \tag{6.1.20}$$

Una vez el tamaño de la población es estimado, es posible utilizar la razón de Hájek (Hájek 1971) para estimar la media poblacional de la siguiente manera.

$$\widetilde{y}_S = \frac{\hat{t}_{y,\pi}}{\hat{N}_{\pi}} \tag{6.1.21}$$

$$= \frac{\sum_{i \in S_I}\frac{t_{yi}}{\pi_{Ii}}}{\sum_{i \in S_I}\frac{N_i}{\pi_{Ii}}} \tag{6.1.22}$$

De hecho, en algunas ocasiones, cuando el diseño de muestreo utilizado induce probabilidades de inclusión desiguales, es mejor utilizar este estimador aun conociendo el tamaño poblacional.

6.1.2 El estimador de Hansen-Hurwitz

Si la selección de los conglomerados se hace con reemplazo, ya sea utilizando un diseño de muestreo aleatorio simple con reemplazo o, en el caso de tener información auxiliar continua a nivel de los conglomerados, haciendo uso de un diseño de muestreo PPT, es posible utilizar los principios del estimador de Hansen-Hurwitz para completar la estrategia de muestreo.

En caso de tener acceso a información auxiliar continua, las probabilidad de selección del i-ésimo conglomerado estaría dada por

$$p_{Ii} = \frac{t_{xi}}{t_x} \tag{6.1.23}$$

Sampath (2001) afirma que en caso de conocerse los tamaños N_i de cada cluster $i = 1, \ldots, N_I$, estos mismos pueden ser utilizados como medidas de tamaño para desarrollar un plan de muestreo con probabilidades proporcionales. El esquema general del muestreo con reemplazo toma la siguiente forma:

- Para cada conglomerado de la población U_I, existen números positivos $p_{I1}, \ldots, p_{IN_I}$ tales que
$$\sum_{U_I} p_{Ii} = 1.$$
Estas probabilidades no son necesariamente iguales.
- Para seleccionar el primer elemento que pertenecerá a la muestra de tamaño m_I, se lleva a cabo un sorteo aleatorio de tal forma que
$$Pr(\text{Seleccionar el conglomerado } i) = p_{Ii}, \ \ i \in U_I.$$
- El conglomerado seleccionado es reemplazado en la población y vuelve a ser parte del próximo sorteo aleatorio con la misma probabilidad de selección. En total se realizan m_I sorteos aleatorios independientes.

Nótese que el sorteo aleatorio se realiza entre los conglomerados, y no entre los elementos; por lo tanto, bajo muestreo en conglomerados no tiene sentido hablar de la probabilidad de selección de un elemento. Una vez que las probabilidades de selección de los conglomerados están definidas, utilizamos el estimador de Hansen-Hurwitz para estimar los parámetros de interés.

Resultado 6.1.6. *Bajo un diseño de muestreo por conglomerados, el estimador de Hansen-Hurwitz para el total t_y, su varianza y su varianza estimada están dados por*

$$\hat{t}_{y,p} = \frac{1}{m_I}\sum_{v=1}^{m_I}\frac{t_{yi_v}}{p_{Ii_v}} \tag{6.1.24}$$

$$Var(\hat{t}_{y,p}) = \frac{1}{m_I}\sum_{i=1}^{N_I} p_{Ii}\left(\frac{t_{yi}}{p_{Ii}} - t_y\right)^2 \tag{6.1.25}$$

$$\widehat{Var}(\hat{t}_{y,p}) = \frac{1}{m_I(m_I-1)}\sum_{v=1}^{m_I}\left(\frac{t_{yi_v}}{p_{Ii_v}} - \hat{t}_{y,p}\right)^2 \tag{6.1.26}$$

respectivamente. Nótese que $\hat{t}_{y,p}$ es insesgado para t_y y que $\widehat{Var}(\hat{t}_{y,p})$ es insesgado para $Var(\hat{t}_{y,p})$.

Prueba. La demostración del resultado sigue los mismos argumentos de la sección del estimador de Hansen-Hurwitz del segundo capítulo y del resultado 2.2.11, definiendo la variable aleatoria Z_v como

$$Z_v = t_{yi}/p_{Ii} \qquad i \in U_I \quad v = 1, \ldots, m_I \tag{6.1.27}$$

y notando que

$$Pr\left(Z_v = t_{yi}/p_{Ii}\right) = p_{Ii} \tag{6.1.28}$$

■

Cochran (1977) afirma que el método de selección de muestras con reemplazo es equivalente al problema estándar de probabilidad en el cual m_I bolas son depositadas en N_I cajas, la probabilidad de que una bola sea depositada en la i-ésima caja está dada por Z_v en cada oportunidad. De esta manera, la distribución conjunta de $n_{Ii}(s_I)$[5] está dada por una expresión multinomial.

Definición 6.1.2. *De manera general, un diseño de muestreo con reemplazo de conglomerados se define como*

$$p_I(s_I) = \begin{cases} \frac{m_I!}{n_{I1}(s_I)!\ldots n_{IN_I}(s_I)!}\prod_{U_I}(p_{Ii})^{n_{Ii}(s_I)}, & si \sum_{U_I} n_{Ii}(s_I) = m_I \\ 0, & en\ otro\ caso \end{cases} \tag{6.1.29}$$

Ejemplo 6.1.3. Nuestra población ejemplo U_I dada por

$$U_I = \{U_1, U_2, U_3\}$$

Suponga que se selecciona una muestra s_I con reemplazo de conglomerados de tamaño $m_I = 2$ mediante un diseño de muestreo que asigna las siguientes probabilidades de selección a cada conglomerado.

$$p_{Ii} = \begin{cases} 0.80, & \text{si } i = 1, \\ 0.15, & \text{si } i = 2, \\ 0.05, & \text{si } i = 3. \end{cases}$$

Para seleccionar una muestra con reemplazo de la población U_I de tamaño $mI = 2$ conglomerados se utiliza la función `sample` cuyo argumento `replace` debe aparecer igual a `TRUE`. Para esto definimos las probabilidades de selección de cada conglomerado.

```
> UI <- c("U1","U2","U3")
> NI=3
> mI=2

> pIi  <- c(0.8, 0.15, 0.05)
> sam  <-  sample(NI, mI, replace=TRUE, prob=pIi)
> mI   <-  UI[sam]
> mI
[1] "U1" "U3"
```

En este caso particular la muestra con reemplazo está compuesta por U_3 y, como era de esperarse, por tener la más alta probabilidad de selección, por U_1. Para estimar el total poblacional, utilizamos la función `HH` del paquete muestreo con los totales de los conglomerados seleccionados y sus respectivas probabilidades de selección.

[5] $n_{Ii}(s_I)$ se define como el número de veces que el conglomerado i-ésimo es seleccionado en una muestra probabilística de tamaño m_I. Note que $n_{Ii}(s_I)$ toma valores $0, 1, 2, \ldots, m_I$.

```
> tyim  <-  tyI[sam]
> tyim
[1] 66 35
> pIim  <-  pIi[sam]

> data.frame(mI,pIim,tyim)
  mI pIim tyim
1 U1 0.80   66
2 U3 0.05   35

> HH(tyim,pIim)[1]
       [,1]
[1,] 391.25
```

6.2 Muestreo aleatorio simple de conglomerados

En esta sección se introducen los principios del diseño de muestreo por conglomerados bajo el plan de muestreo más sencillo. La muestra s_I de n_I conglomerados es seleccionada mediante un diseño de muestreo aleatorio simple sin reemplazo. Como se verá a lo largo de la sección, no hay nuevos principios (ni en el diseño de muestreo ni en el desarrollo del estimador) involucrados en la construcción de la estrategia de muestreo, la demostración de los resultados se hace siguiendo las pautas expuestas en el capítulo 2.

Este diseño de muestreo asume que el comportamiento del total de la característica de interés es constante en cada uno de los conglomerados. En la práctica esta situación se presenta en muy pocas ocasiones, es por esto que este diseño pierde precisión, en la mayoría de ocasiones, ante el muestreo aleatorio simple. Para que este diseño de muestreo sea más eficiente el valor promedio de la característica de interés en cada cluster $\bar{y}_{U_i}$ debería ser proporcional a $\frac{c}{N_i}$. Se asume que la población U_I está dividida en N_I conglomerados (no necesariamente del mismo tamaño). La muestra sin reemplazo es seleccionada de acuerdo al diseño de muestreo dada en la siguiente definición.

Definición 6.2.1. *Un diseño de muestreo se dice aleatorio simple para conglomerados si todas las posibles muestras de tamaño n_I tienen la misma probabilidad de ser seleccionadas. Así,*

$$p_I(s_I) = \begin{cases} \frac{1}{\binom{N_I}{n_I}} & \text{si } \#s_I = n_I \\ 0 & \text{en otro caso} \end{cases} \tag{6.2.1}$$

Una vez que la muestra de conglomerados s_I es seleccionada se dispone a realizar una enumeración completa y la respectiva medición y observación de todos y cada uno de los elementos pertenecientes a cada conglomerado seleccionado.

6.2.1 Algoritmos de selección

En la selección de las muestras de conglomerados sin reemplazo es posible utilizar los algoritmos de muestreo dados en el capítulo 2, de tal forma que los siguientes

pasos se deben realizar:

- Separar la población en N_I conglomerados mediante el marco de muestreo de conglomerados.
- Realizar una selección de n_I conglomerados mediante cualquiera de los métodos expuestos en la sección 3.2.1; es decir, por el método coordinado negativo o por el método de Fan-Muller-Rezucha.

6.2.2 El estimador de Horvitz-Thompson

Siguiendo el resultado 6.1.1. las probabilidades de inclusión están dadas por el siguiente resultado.

Resultado 6.2.1. *Para un diseño de muestreo aleatorio de conglomerados, las probabilidades de inclusión de primer y segundo orden de los conglomerados están dadas por*

$$\pi_{Ii} = \frac{n_I}{N_I} \tag{6.2.2}$$

$$\pi_{Iij} = \frac{n_I(n_I-1)}{N_I(N_I-1)} \tag{6.2.3}$$

respectivamente.

Resultado 6.2.2. *El tamaño de la muestra de elemento s es aleatorio y su esperanza está dada por*

$$E(n(S)) = N\frac{n_I}{N_I} \tag{6.2.4}$$

Prueba. De la definición de tamaño de muestra esperado, se tiene que

$$E(n(S)) = E\left(\sum_{i\in S_I} N_i\right) = \sum_{i\in U_I} N_i\frac{n_I}{N_I} = N\frac{n_I}{N_I} \tag{6.2.5}$$

■

Se sigue del resultado 6.1.2 que la estrategia de muestreo se construye mediante el uso del estimador de Horvitz-Thompson que bajo este diseño de muestreo particular toma la forma del siguiente resultado.

Resultado 6.2.3. *Para un diseño de muestreo aleatorio de conglomerados, el estimador de Horvitz-Thompson del total poblacional t_y, su varianza y su varianza estimada están dados por*

$$\hat{t}_{y,\pi} = \frac{N_I}{n_I}\sum_{S_I} t_{yi} \tag{6.2.6}$$

$$Var_{MAC}(\hat{t}_{y,\pi}) = \frac{N_I^2}{n_I}\left(1-\frac{n_I}{N_I}\right)S^2_{t_{yU_I}} \tag{6.2.7}$$

$$\widehat{Var}_{MAC}(\hat{t}_{y,\pi}) = \frac{N_I^2}{n_I}\left(1-\frac{n_I}{N_I}\right)S^2_{t_{ys_I}} \tag{6.2.8}$$

respectivamente, con $S^2_{t_{yU_I}}$ *y* $S^2_{t_{ys_I}}$ *el estimador de la varianza de los totales de los conglomerados para la característica de interés en el universo* U_I *y en la muestra* s_I*. Esto es*

$$S^2_{t_{yU_I}} = \frac{1}{N_I - 1}\sum_{i\in U_I}(t_{yi} - \bar{t}_{U_I})^2,$$

donde $\bar{t}_{U_I} = \sum_{i=1}^{N_I} t_{yi}/N_I$*, y* $S^2_{t_{yS_I}}$ *se define de manera análoga. Nótese que* $\hat{t}_{y,\pi}$ *es insesgado para el total poblacional* t_y *de la característica de interés* y*, y que* $\widehat{Var}_{MAC}(\hat{t}_{y,\pi})$ *es insesgado para* $Var_{MAC}(\hat{t}_{y,\pi})$*.*

Nótese que el diseño de muestreo sistemático en un caso especial del muestreo aleatorio de conglomerados cuando se selecciona una muestra s_I de tamaño igual a $n_I = 1$. Al igual que en muestreo sistemático no se tiene un estimador de la varianza cuando se selecciona sólo un conglomerado.

Ejemplo 6.2.1. Siguiendo con nuestra población ejemplo U_I, existen $\binom{N_I}{m_I} = \binom{3}{2} = 3$ posibles muestras de tamaño $m_I = 2$. Realice el cálculo léxico-gráfico del estimador de Horvitz-Thompson y compruebe el insesgamiento y la varianza mediante este diseño de muestreo.

Tamaño de muestra

Bajo muestreo aleatorio de conglomerados se utilizan los mismos principios de la estimación del tamaño de muestra en muestreo aleatorio simple reemplazando las cantidades correspondientes de la población de elementos por la población de conglomerados U_I. De tal forma que si requiere estimar el tamaño de muestra dada una precisión absoluta c se tiene:

$$n_I \geq \frac{n_{I0}}{1 + \frac{n_{I0}}{N_I}} \tag{6.2.9}$$

con $n_{I0} = \dfrac{t^2_{1-\alpha/2,N_I-1}S^2_{t_{yU_I}}}{c^2}$. En algunas ocasiones se quiere lograr una precisión relativa k, por tanto:

$$n_I \geq \frac{n_{I0}}{1 + \frac{n_{I0}}{N_I}} \tag{6.2.10}$$

con $n_{I0} = \dfrac{t^2_{1-\alpha/2,N_I-1}CV^2}{k^2}$. Nótese que dado que la población de conglomerados es pequeña, en la mayoría de los casos, es preferible suponer que el estimador sigue una distribución t-student con $N_I - 1$ grados de libertad.

6.2.3 Eficiencia de la estrategia

A lo largo del capítulo se ha mencionado que la eficiencia de esta estrategia de muestreo es menor que la del muestreo aleatorio simple sin reemplazo. Intuitivamente se sospecha que, dado que la formación de grupos se presenta en forma

natural en la mayoría de los casos, la información de los conglomerados, con respecto al comportamiento estructural de la característica de interés, es homogénea dentro de cada uno de ellos.

Para corroborar las anteriores afirmaciones vamos a medir la eficiencia de la estrategia utilizando el efecto de diseño. Sin embargo, para unificar el tamaño de la muestra en esta estrategia se supondrá que:

1. La población U_I está conformada por N_I conglomerados.
2. Cada conglomerado es de tamaño M. Luego $\#U_i = M \quad i = 1, \ldots, N_I$, además la población de elementos U es de tamaño $N = M \times N_I$.
3. Se selecciona una muestra s_I de tamaño igual a n_I conglomerados. De esta forma se han seleccionado en la muestra $M \times n_I$ elementos.

Tabla 6.1: *Tabla de ANOVA inducida por el muestreo aleatorio de conglomerados.*

Fuente	gl	Suma de cuadrados	Cuadrado medio
Entre	$N_I - 1$	$SCE = \sum_{i=1}^{N_I} M\left(\bar{y}_{U_i} - \bar{y}_U\right)^2$	$\dfrac{SCE}{N_I - 1}$
Dentro	$N_I M - N_I$	$SCD = \sum_{i=1}^{N_I}\sum_{j=1}^{M}\left(y_{ij} - \bar{y}_{U_i}\right)^2$	$\dfrac{SCD}{N_I M - N_I}$
Total	$N_I M - 1$	$SCT = \sum_{i=1}^{N_I}\sum_{j=1}^{M}\left(y_{ij} - \bar{y}_U\right)^2$	$s^2_{y_U}$

Los resultados podrán ser comparables si se supone que una muestra de n_I conglomerados es seleccionada de acuerdo a un diseño aleatorio simple de conglomerados. Por otro lado, se supone que se selecciona una muestra de $M \times n_I$ elementos directamente de la población U. Cada vez que la población es dividida en sub-grupos poblacionales es muy útil recurrir a la tabla de análisis de varianza que esta vez toma la forma dada en la tabla 6.1.

Resultado 6.2.4. *Utilizando los resultados de la descomposición de las sumas de cuadrados, la varianza de la estrategia por conglomerados toma la siguiente forma*

$$Var_{MAC}(\hat{t}_{y,\pi}) = \frac{N_I^2}{n_I}\left(1 - \frac{n_I}{N_I}\right) M \frac{SCE}{N_I - 1} \tag{6.2.11}$$

mientras que la varianza de la estrategia aleatoria simple, con un tamaño poblacional igual a $N = M \times N_I$ elementos y un tamaño de muestra igual a $n = M \times n_I$ elementos, se puede escribir como

$$Var_{MAS}(\hat{t}_{y,\pi}) = \frac{N_I^2}{n_I}\left(1 - \frac{n_I}{N_I}\right) M \frac{SCT}{MN_I - 1} \tag{6.2.12}$$

Prueba. Para la varianza de la estrategia por conglomerados se tiene

$$\begin{aligned} M\frac{SCE}{N_I - 1} &= \frac{\sum_{i=1}^{N_I} M^2\left(\bar{y}_{U_i} - \bar{y}_U\right)^2}{N_I - 1} \\ &= \frac{\sum_{i=1}^{N_I}\left(t_{yi} - \bar{t}_{yU_I}\right)^2}{N_I - 1} \\ &= S^2_{t_{yU_I}} \end{aligned}$$

donde $\bar{y}_{U_i}$ y t_{yi} es el promedio y el total del i-ésimo conglomerado, respectivamente y $\bar{t}_{yU_I} = \frac{\sum_{i=1}^{N_I} t_{yi}}{N_I}$ es el promedio de los totales de los conglomerados.
Para la varianza de la estrategia aleatoria simple sólo hay que notar que

$$\begin{aligned}\frac{N^2}{n}\left(1-\frac{n}{N}\right) &= \frac{(MN_I)^2}{Mn_I}\left(1-\frac{Mn_I}{MN_I}\right)\\ &= \frac{N_I^2}{n_I}\left(1-\frac{n_I}{N_I}\right)M\end{aligned}$$

■

Note que si SCE es alta, entonces la estrategia será menos eficiente. En la práctica es esto lo que ocurre exactamente pues dada la agrupación natural de elementos, el comportamiento de la característica de interés será similar dentro de cada conglomerado. Por tanto, SCE será elevada pues en forma general los conglomerados presentarán un comportamiento heterogéneo. Para verlo con más claridad, se define el coeficiente de correlación intra-clase como

$$\rho = 1 - \frac{M}{M-1}\frac{SCD}{SCT} \tag{6.2.13}$$

Esta medida toma valores positivos si los elementos dentro de los conglomerados tienen un comportamiento similar y negativo cuando el comportamiento de los elementos dentro de los conglomerados es muy disperso. Además, el coeficiente informa qué tan similares son los elementos dentro de los conglomerados y proporciona una medida de homogeneidad dentro de los conglomerados y nos da una realidad más detallada en cuanto al efecto de diseño y pérdida de eficiencia en el muestreo aleatorio de conglomerados como lo afirma el siguiente resultado.

Resultado 6.2.5. *El efecto de diseño en muestreo aleatorio de conglomerados está dado por*

$$Deff = \frac{Var_{MAC}\hat{t}_\pi}{Var_{MAS}\hat{t}_\pi} \cong 1 + (M-1)\rho \tag{6.2.14}$$

Prueba. La aproximación se tiene si se supone que N_I, el número total de conglomerados, es grande tal que

$$M(N_I - 1) \cong MN_I - 1 \tag{6.2.15}$$

La demostración se completa notando que al realizar el cociente de varianzas, al igual que en la sección del muestreo sistemático, se tiene que

$$\frac{SCE}{SCT} = \frac{1+(M-1)\rho}{M} \tag{6.2.16}$$

■

Dado que ρ es generalmente positivo[6] podemos inferir de (6.2.14) que el muestreo por conglomerados tendrá una mayor varianza que el muestreo aleatorio simple

[6]Esto se da porque los conglomerados se forman física y geográficamente como agrupaciones contiguas de elementos que comparten un ambiente natural, entonces el comportamiento de los elementos internamente será similar.

de elementos directamente de la población U. Sin embargo, es plausible sacrificar la eficiencia estadística por el ahorro financiero y logístico característico de las estrategias por conglomerados. Ahora, si ρ es negativo, esta estrategia gana en eficiencia y también en costos operativos.

Lohr (2000) afirma que en el caso, muy común en la práctica, en que los conglomerados no sean del mismo tamaño, una medida alternativa a ρ es el coeficiente de determinación R^2 definido como

$$R^2 = 1 - \frac{CMD}{s^2_{y_U}} \tag{6.2.17}$$

donde $CMD = \frac{SCD}{N-N_I}$; con N el número total de elementos en la población U. Ésta es una medida muy conocida y utilizada en el análisis de regresión lineal, y es interpretada como la cantidad de variabilidad explicada por los promedios de cada conglomerado. Si el comportamiento de la característica de interés es homogéneo dentro de los conglomerados, entonces los promedios entre los conglomerados tendrán una muy alta dispersión con respecto a la variación dentro de los conglomerados y R^2 tomará valores grandes.

6.2.4 Marco I y Lucy

El común denominador de las aplicaciones prácticas con Marco y Lucy en los capítulos anteriores ha sido la identificación y ubicación, a priori, de cada una de las empresas en el sector industrial. Esto ha sido posible gracias a que un marco de muestreo de elementos estuvo disponible. En algunas ocasiones, el marco de muestreo disponible mostró bondades que permitieron la incorporación de información auxiliar, ya sea de tipo continuo o categórico, para mejorar la eficiencia de la estrategia de muestreo utilizada en cada caso.

En cualquier caso, el gobierno desea obtener estimaciones precisas que le permitan fortalecer sus políticas de apoyo y financiamiento de las empresas en el sector industrial. Sin embargo, el gobierno no está en disposición de entregar una lista de todas las empresas del sector industrial con su respectiva identificación y ubicación debido a políticas de confidencialidad que no le permiten brindar este tipo de información. Por tanto, en esta ocasión no hay tal marco generoso de elementos en la población y el estudio se deberá llevar a cabo con esta restricción de tipo logístico.

En cualquier estudio por muestreo, siempre debe existir, si no físicamente al menos de forma implícita, un marco de muestreo de la población que permita llegar a la medición de la unidad objetivo de muestreo. Dado que el gobierno no permite la utilización de un marco de muestreo de empresas en el sector industrial, se debe realizar el levantamiento de un marco de muestreo de conglomerados que agrupen estas empresas. Una solución, que es muy utilizada en la práctica, es realizar un muestreo de áreas geográficas. Las empresas, las viviendas, los domicilios, los negocios, etc. están ubicadas en algún lugar del mapa y es poco factible que se muevan de donde han estado instaladas. Por tanto, un marco de muestreo por áreas es una buena solución de tipo logístico para enfrentar la etapa de diseño de este estudio.

Un inconveniente que se presenta a la hora de realizar un muestreo de conglomerados con un marco discriminado en áreas geográficas es la imposibilidad de conocer cuántas empresas estarán ubicadas en cada zona geográfica. Sin embargo, sí es posible asignar subdivisiones de cada zona geográfica seleccionada a un grupo de encuestadores para que recorran la zona y apliquen el cuestionario a cada una de las empresas del sector. De esta forma, es posible tener una estimación del presupuesto que se requiere. La población U_I de conglomerados, es decir la ciudad, se divide en cinco zonas geográficas, a saber: **Zona A**, ubicada en el sur, **Zona B**, ubicada en el norte, **Zona C**, ubicada en el oriente, **Zona D**, ubicada en el occidente y **Zona E**, ubicada en el centro.

Recordando los objetivos del estudio, el gobierno quiere medir el crecimiento del sector industrial en la ciudad, mediante tres características importantes: el ingreso y los impuestos declarados en el último año fiscal y la generación de empleos mediante la cantidad de trabajadores que laboran en cada empresa. Seguramente, ni el ingreso, ni los impuestos, ni la cantidad de empleados están correlacionados con la zona geográfica. Podemos afirmar esto porque la ubicación de las empresas es realizada por el gobierno siguiendo diversos criterios.

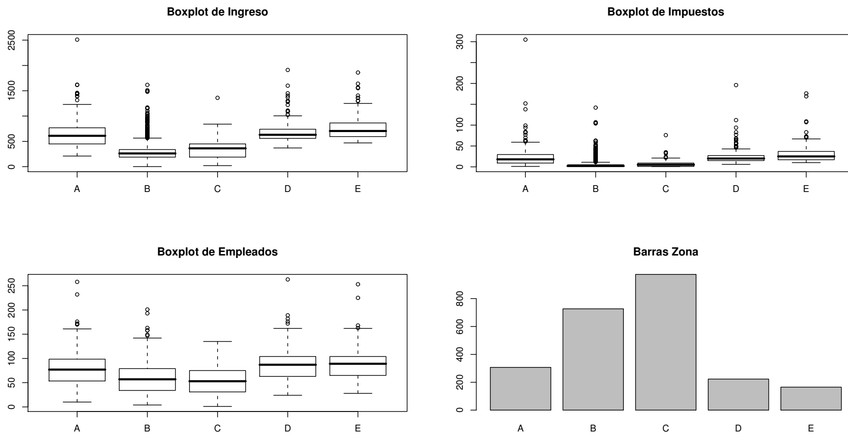


Figura 6.1: *Boxplot de las características de interés en cada zona geográfica.*

Es así como en una misma zona geográfica, es posible encontrar una empresa grande rodeada de empresas pequeñas o medianas. Este es un muy buen indicio en la etapa del diseño de muestreo pues quiere decir que el comportamiento de las características de interés dentro de cada área geográfica es muy disperso. La figura 6.1 presenta el comportamiento de las características de interés en cada una de las cinco zonas geográficas de la ciudad. Nótese que no es posible identificar un comportamiento estructural significativamente diferente en cada zona, sino que por el contrario, el comportamiento es heterogéneo dentro de cada zona y homogéneo entre las zonas.

Aunque no se conoce el número de empresas en el sector industrial, el gobierno ha estimado según datos de años anteriores la existencia de 2500 empresas para el último año fiscal. Con esta información se ha decidido seleccionar una mues-

tra aleatoria simple de conglomerados de tamaño $n_I = 2$. Por tanto, el tamaño muestral de empresas esperado corresponde a $2500\frac{2}{5} = 1000$. De la población de $N_I = 5$ conglomerados de áreas se selecciona una muestra aleatoria simple de $n_I = 2$ utilizando la función `S.SI` del paquete muestreo. En este caso particular, los conglomerados incluidos en la muestra sin reemplazo corresponde a la **Zona A** y a la **Zona E**.

```
> UI <- c("A","B","C","D","E")
> NI=length(UI)
> nI=2

> samI <- S.SI(NI,nI)
> muestra <- UI[samI]
> muestra
[1] "A" "E"
```

Un equipo de encuestadores se dispone a recolectar la información de cada una de las empresas pertenecientes a los conglomerados seleccionados, el plan operativo es más eficiente entre más entrevistadores sean contratados por cada conglomerado seleccionado. Cuando el proceso de medición termina se tienen dos conjuntos de datos, cada una conteniendo el valor de las características de interés para cada una de las empresas del área, correspondientes a **Zona A** y **Zona E**.

Con la función `rbind` es posible unir la información de las zonas geográficas seleccionadas en la muestra. Con ayuda de la función `T.SIC(y,C)`, del paquete `TeachingSampling`, es posible obtener los totales de las características de interés en cada conglomerado. Los argumentos de esta función son `y`, el conjunto de datos (una sola variable o un conjunto de variables) del censo en cada conglomerado y `C`, una variable que indica la pertenencia del elemento, en este caso de las empresas, al conglomerado. El resultado de la función es el total de elementos en cada conglomerado, así como el total de las características de interés en cada uno de los conglomerados. En este caso particular, el tamaño de la muestra de empresas es $307 + 165 = 472$. Nótese que, como en los casos de estimación de los capítulos anteriores, se crea un conjunto de datos de las características de interés definido por `estima <- data.frame(Income, Employees, Taxes)`.

```
> data(Lucy)
> attach(Lucy)
> Lucy1 <- Lucy[which(Zone==muestra[1]),]
> Lucy2 <- Lucy[which(Zone==muestra[2]),]

> LucyI <- rbind(Lucy1,Lucy2)
> attach(LucyI)

> Area <- as.factor(as.integer(Zone))
> estima <- data.frame(Income, Employees, Taxes)
> y <- T.SIC(estima,Area)
```

```
> y
   1 Income Employees Taxes
1 307  200251     24338    7377.5
5 165  126619     14570    5131.0
```

Una vez que se tienen los totales de cada zona geográfica, se utiliza la función `E.SI(NI,nI,y)` del paquete muestreo, definida en el capítulo dos, para obtener las estimaciones de los parámetros de interés.

```
> E.SI(NI,nI,y)
                                  Income    Employees    Taxes
Total estimado           8.171750e+05 9.727000e+04 3.127125e+04
Varianza                 2.033127e+10 3.578018e+08 1.892536e+07
coeficiente de variación 1.744886e+01 1.944654e+01 1.391159e+01
```

Los resultados de la estimación se muestran en la siguiente tabla. Es de considerar que la eficiencia de esta estrategia de muestreo es mucho menor que la de una estrategia que utilice un diseño de muestreo aleatorio simple. Nótese que la desviación relativa es mucho mayor.

Tabla 6.2: *Muestreo aleatorio de conglomerados: estimación de los totales de las características de interés.*

Variable	Total poblacional	Total estimado	cve %	Desv. %
Ingreso	1035217	817175	1.74	-21.06
Empleados	151950	97270	1.94	-35.99
Impuestos	28654	31271	1.39	9.14

Es claro que los resultados de esta estrategia de muestreo no son satisfactorios, por lo menos para la estimación de los parámetros de interés de Ingreso e Impuestos. La explicación de la deficiencia de esta estrategia es inmediata al analizar el siguiente gráfico que muestra el comportamiento estructural de los totales en los conglomerados.

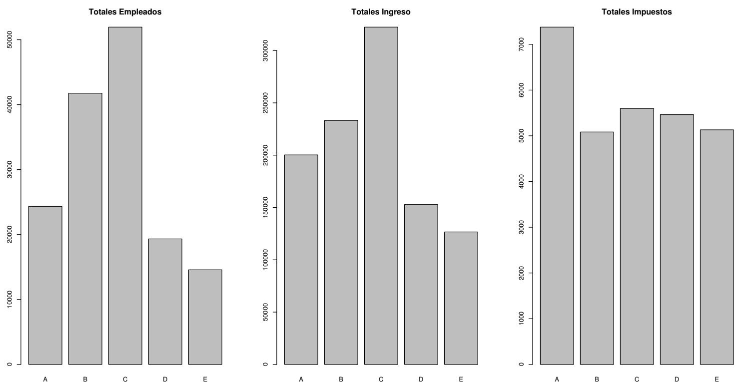


Figura 6.2: *Totales de las características de interés en cada conglomerado.*

Es notable como el comportamiento de los totales es tan diferente en cada conglomerado en las características Ingreso y Empleados. Sin embargo, el comportamiento es similar en cuanto a la característica Impuestos. Es interesante observar que entre más disimilitud existe entre los totales de los conglomerados, más desviación relativa existe en la estimación. Como se dijo en la introducción de este capítulo, esta estrategia de muestreo es ineficiente en aquellos casos en que los totales de cada conglomerado no están correlacionados con las probabilidades de inclusión a nivel de conglomerados. Observando el gráfico, se establece que Impuestos es la única característica que presenta un comportamiento estable en relación a los conglomerados.

La regla de oro se mantiene, una estrategia de muestreo es eficiente si las probabilidades de inclusión están correlacionadas con los valores de la característica de interés, en este caso con los totales de cada conglomerados.

6.3 Ejercicios

6.1 Argumente si las siguientes afirmaciones son falsas o verderas. Sustente su respuesta detallamente.

(a) En un diseño de muestreo de conglomerados, siempre se tiene un marco de muestreo de elementos de la población.

(b) En un diseño de muestreo de conglomerados, para la estimación de un total, se obtiene mayor precision si las probabilidades de selección o inclusión son proporcionales a los totales de la característica de interés en los conglomerados.

(c) En un diseño de muestreo de conglomerados, para la estimación de un total, se obtiene mayor precision si las probabilidades de selección o inclusión son proporcionales a la característica de interés de los elementos en los conglomerados.

(d) En la estimación de totales poblaciones, se nota que, casi siempre, $Var_{MASC}(t_{y,\pi})$ es mayor a $Var_{MAS}(t_{y,\pi})$.

(e) En un diseño de muestreo aleatorio simple de conglomerados de tamaño desigual, hay un aumento significativo de la varianza, respecto a un diseño de muestreo aleatorio simple de conglomerados de igual tamaño.

(f) En un diseño de muestreo PPT de conglomerados de tamaño desigual (con probabilidad proporcional al tamaño del conglomerado), hay una disminución significativa de la varianza, respecto a un diseño de muestreo aleatorio simple de conglomerados de tamaño desigual.

6.2 Suponga que el objetivo de una encuesta es estimar el ingreso medio en un barrio de la ciudad. Asuma que en ese barrio existen $N_I = 60$ manzanas. Se realiza un diseño de muestreo aleatorio simple de conglomerados y se seleccionan $n_I = 5$ manzanas, en las cuales se entrevistan a todos los hogares. Los resultados de la encuesta se dan en la tabla 6.3

(a) Estime el ingreso total de los hogares en el barrio. Reporte el coeficiente de variación estimado.

Tabla 6.3: Tabla de las cinco manzanas seleccionadas: ejercicio 6.2

ID Manzana	Hogares en la manzana	Ingreso total en la manzana
AW45	120	25000
AW02	100	24000
AW31	80	19000
AW28	95	20100
AW44	80	18000

(b) Estime el número de hogares en el barrio. Reporte el coeficiente de variación estimado.

(c) Asumiendo que en el barrio hay $N = 2000$ hogares, estime el ingreso medio de los hogares en el barrio. Reporte el coeficiente de variación estimado.

(d) Estime el ingreso medio utilizando el estimador de Hájek. Explique la diferencia con respecto a la estimación del punto anterior.

Capítulo 7

Muestreo en varias etapas

> En muchas situaciones, los elementos de un conglomerado pueden ser demasiado similares, de modo que el análisis de todos los elementos que conforman el conglomerado será un desperdicio de recursos. En estos casos podría ser más barato seleccionar más conglomerados y tomar una submuestra dentro de cada uno de ellos.
>
> Sharon Lohr (2000)

En el capítulo anterior se utilizó la agrupación natural de los elementos en la población para ahorrar costes financieros y logísticos al planear una estrategia de muestreo por conglomerados. Sin embargo, el ahorro en términos operativos se ve reflejado en un alto precio por pagar con respecto a la eficiencia estadística de la estrategia. Una posible solución para disminuir la varianza es aumentar el tamaño de muestra de conglomerados, solución que aumentaría los costos operativos.

Para mantener un equilibrio entre los costos financieros y las bondades de la estrategia de muestreo es posible aprovechar la homogeneidad dentro de los conglomerados y, de esta manera, no realizar un censo dentro de cada conglomerado seleccionado sino proceder a seleccionar una sub-muestra dentro del conglomerado seleccionado. Como el comportamiento estructural de la característica de interés al interior de los conglomerados es homogéneo, entonces una estimación del total del conglomerado tendría una varianza pequeña. Por supuesto, como no se tienen acceso a un marco de muestreo de elementos, se debe realizar un empadronamiento para levantar un marco de muestreo de elementos en cada uno y sólo en los conglomerados seleccionados. Una vez se disponga del marco de muestreo de elementos dentro de los conglomerados, se dispone la selección de las sub-muestras de elementos. Bautista (1998) plantea que el principio básico del muestreo en varias etapas se puede definir como el proceso jerárquico que realiza l veces los siguientes pasos:

1. Construcción de l marcos de muestreo de unidades (conglomerados en las primeras $l-1$ etapas del diseño muestral y de elementos en la última etapa).
2. Aplicación de un diseño muestral y selección de la muestras (o sub-muestras) de cada marco de muestreo.

Nótese que se ha introducido el concepto de **unidad de muestreo** refiriéndose a conglomerados de elementos o a los elementos. Si el diseño de muestreo tiene tres etapas, por ejemplo: si se quieren obtener estimaciones acerca del comportamiento de los alumnos en determinada ciudad, y no se dispone de un marco de muestreo de los alumnos, es posible en una **primera etapa** levantar un marco de muestreo de todas y cada una de las escuelas en la ciudad y realizar una selección de una muestra de escuelas mediante cierto diseño de muestreo. Una vez que las escuelas son seleccionadas, en una **segunda etapa**, se levanta un marco de muestreo de niveles académicos dentro de las escuelas (cursos o clases) y se procede a seleccionar una muestra de niveles. De tal forma que en la **tercera y última etapa** se levanta un marco de muestreo de elementos; es decir, de alumnos pertenecientes a cada nivel seleccionado, y se realiza una muestra de elementos que serán observados y medidos.

Es interesante observar cómo la población, en el estado de la naturaleza, se subdivide gracias al comportamiento «jerárquico», que en este caso particular toma la siguiente forma:

$$\underbrace{\textbf{Ciudad}}_{\text{Población}U} \Rrightarrow \underbrace{\textbf{Escuelas}}_{\text{UPM}} \Rrightarrow \underbrace{\textbf{Niveles}}_{\text{USM}} \Rrightarrow \underbrace{\textbf{Alumnos}}_{\text{UTM}}$$

Como notación, se llama **Unidad Primaria de Muestreo** o **UPM** a la primera subdivisión en conglomerados de la población original, **Unidad Secundaria de Muestreo** o **USM** a la sub-subdivisión de la población, es decir la subdivisión de las **UPM**. La **Unidad Terciaria de Muestreo** o **UTM** corresponde a los elementos de la población objetivo, que en este caso particular son los alumnos de la ciudad.

No siempre las unidades finales de muestreo son elementos, es así como es posible planear un diseño en dos etapas de conglomerados, refiriéndose a que la unidad secundaria de muestreo son conglomerados, o también es posible aplicar un diseño en cuatro etapas de elementos, en donde las unidades finales de muestreo sean elementos; por ejemplo, en Bautista (1998) se presenta el siguiente caso:

$$\underbrace{\textbf{Ciudad}}_{\text{Población}U} \Rrightarrow \underbrace{\textbf{Sección}}_{\text{UPM}} \Rrightarrow \underbrace{\textbf{Manzana}}_{\text{USM}} \Rrightarrow \underbrace{\textbf{Vivienda}}_{\text{UTM}} \Rrightarrow \underbrace{\textbf{Persona}}_{\text{UCM}}$$

El principio básico de una estrategia de muestreo en varias etapas es construir estimaciones desde abajo hasta arriba. Pero para que los resultados de la estimación basada en el diseño de muestreo sean aplicables, se deben satisfacer los siguientes dos supuestos:

1. **Invariancia:** sugiere que la probabilidad de selección de una muestra de unidades de muestreo (conglomerados o elementos) no depende del diseño de muestreo de la anterior etapa.
2. **Independencia:** interpretado como que el sub-muestreo de cualquier unidad de muestreo se lleva a cabo de manera independiente con las otras unidades de muestreo, en la misma etapa o en etapas superiores o inferiores.

Para el resto del capítulo se asume implícitamente que estas propiedades se satisfacen en cada etapa de muestreo de la estrategia. Si los supuestos no se satisfacen,

entonces el lector puede consultar la sección de muestreo en varias fases del capítulo de Tópicos avanzados. Para asentar aún más la filosofía interna del muestreo en varias etapas, es necesario estudiar el más simple de todos los diseños de muestreo de esta clase: el muestreo en dos etapas.

7.1 Muestreo en dos etapas

También llamado muestreo «bietápico» por Mahalanobis (1946), este diseño de muestreo estima el total de cada cluster t_i mediante una sub-muestra dentro de los conglomerados seleccionados de la población. En la estimación de los parámetros de interés se encuentran dos fuentes de variabilidad cada una en cada etapa. Es decir, existe variabilidad debido a la selección de las unidades primarias de muestreo o conglomerados y, por supuesto, también existe variabilidad debido a la selección de una muestra de elementos, unidades secundarias de muestro en los conglomerados seleccionados.

Suponga que la población de elementos U se divide en N_I **unidades primarias de muestreo**, que definen una partición de la población, llamados también **conglomerados** y denotadas como $U_I = \{U_1, \ldots, U_{N_I}\}$. El i-ésimo conglomerado U_i $i = 1, \ldots, N_I$ es de tamaño N_i. Särndal, Swensson & Wretman (1992) dan un marco general para el muestreo en dos etapas, de tal manera que

1. Una muestra s_I de unidades primarias de muestreo es seleccionada de U_I de acuerdo a un diseño de muestreo $p_I(s_I)$. Nótese que S_I representa la muestra aleatoria de conglomerados tal que $Pr(S_I = s_I) = p_I(s_I)$.

2. Para cada conglomerado U_i $i = 1, \ldots, N_I$ seleccionado en la muestra s_I, se selecciona una muestra s_i de elementos seleccionada de acuerdo a un diseño de muestreo $p_i(s_i)$. Nótese que S_i representa la muestra aleatoria de elementos tal que $Pr(S_i = s_i) = p_i(s_i)$.

Este diseño de muestreo bietápico debe cumplir las dos propiedades de invarianza y de independencia. La invarianza significa que los diseños de muestreo $p_i(s_i)$ de la segunda etapa **no dependen** del resultado en la primera etapa, es decir, que el diseño de muestreo siempre debe ser el mismo dentro de cada una de las unidades primarias de muestreo.

$$Pr(S_i = s_i \mid S_I = s_I) = Pr(S_i = s_i). \tag{7.1.1}$$

Nótese que lo anterior implica que $p_i(\cdot|s_I) = p_I(\cdot)$

La independencia significa que el proceso de selección de muestras en la segunda etapa dentro de cada unidad primaria de muestreo no depende de los procesos de selección utilizados en los restantes unidades primarias de muestreo. Es decir, el submuestreo en una unidad primaria de muestreo particular es independiente del submuestreo en otras unidades primarias de muestreo [1], por tanto, para cada muestra aleatoria S_I en la primera etapa se cumple que

[1] Nótese el símil con el proceso de estratificación.

$$Pr\left(\bigcup_{i\in s_I} s_i|s_I\right) = \prod_{i\in s_I} Pr(s_i|s_I) \tag{7.1.2}$$

Si el diseño de muestreo en la primera etapa es con reemplazo, entonces un conglomerado puede aparecer más de una vez, y se debe proceder a realizar el submuestreo tantas veces como aparezca dicha unidad primaria en la muestra realizada s_I, con esto se garantiza que se cumplan las propiedades de independencia e invarianza. En términos de soporte, es posible hablar de también del tres clases de soporte. A saber:

- En la primera etapa existe un soporte Q_I conteniendo todas las posibles muestras realizadas de las unidades primarias de muestreo.
- En la segunda etapa existe un soporte Q^i para cada $i \in U_I$, es decir, para cada unidad primaria en la etapa anterior.
- En general, el soporte Q conteniendo todas las posibles muestras de elementos mediante un diseño bietápico está dado por

$$\begin{aligned} Q &= \bigcup_{r=1}^{\#Q_I} \bigcup_{i\in s_I^{(r)}} s_i, \quad \text{con} \quad s_i \in Q^i \\ &= \left\{ \bigcup_{i\in s_I^{(r)}} s_i, \quad \text{con} \quad s_i \in Q^i, r = 1, \ldots, \#Q_I \right\} \end{aligned} \tag{7.1.3}$$

Donde $s_I^{(r)}$ denota la r-ésima posible muestra en la primera etapa y la cardinalidad de Q está dada por

$$\#Q = \prod_{i\in U_I} \#Q^i$$

Y la muestra de elementos - o unidades secundarias de muestreo - viene dada por

$$S = \bigcup_{i\in S_I} S_i, \text{ con } \ S_i \in Q^i \tag{7.1.4}$$

con tamaño de la muestra aleatorio dado por

$$n(S) = \sum_{i\in S_I} n_i \tag{7.1.5}$$

La definición de los soportes en cada etapa y, en general, nos permiten proclamar que el diseño de muestreo bietápico es un auténtico diseño de muestreo.

Resultado 7.1.1. *El diseño de muestreo bietápico cumple que*

1. $p(s) \geq 0$ *para todo* $s \in Q$
2. $\sum_{s\in Q} p(s) = 1$

Prueba. En primer lugar, se tiene que

$$\begin{aligned}
p(s) &= Pr(\text{Seleccionar } s_I \text{ en la etapa uno y seleccionar } \bigcup_{i\in s_I} s_i \text{ en etapa dos})\\
&= \underbrace{p_I(s_I)\, Pr\left(\bigcup_{i\in s_I} s_i|s_I\right)}_{Independencia}\\
&= p_I(s_I)\prod_{i\in s_I}\underbrace{Pr(s_i|s_I)}_{Invarianza}\\
&= p_I(s_I)\prod_{i\in s_I} p_i(s_i)
\end{aligned}$$

y es claro que $p(s) \geq 0$. Ahora, para demostrar la segunda propiedad, se tiene que

$$\begin{aligned}
\sum_{s\in Q} p(s) &= \sum_{r=1}^{\#Q_I}\sum_{s_I^{(r)}} p(s)\\
&= \sum_{r=1}^{\#Q_I}\sum_{s_I^{(r)}} p_I(s_I^{(r)})\prod_{i\in s_I^{(r)}} p_i(s_i)\\
&= \sum_{r=1}^{\#Q_I} p_I(s_I^{(r)})\underbrace{\sum_{s_I^{(r)}}\prod_{i\in s_I^{(r)}} p_i(s_i)}_{=1}\\
&= \sum_{r=1}^{\#Q_I} p_I(s_I^{(r)}) = 1
\end{aligned}$$

En donde la equivalencia a uno del segundo sumando en la tercera igualdad se obtiene haciendo el símil con la demostración del resultado 5.1.1., en donde el diseño estratificado se definió como una productoria. ■

Para ilustrar el anterior resultado, junto con la compenetración de los conceptos de soportes en cada una de las etapas, se diseñó el siguiente ejemplo que utiliza un diseño de muestreo sin reemplazo en dos etapas.

Ejemplo 7.1.1. Nuestra población ejemplo U_I dada por

$$U_I=\{U_1, U_2, U_3\}$$

Suponga que se selecciona una muestra s_I de unidades primarias de muestreo de tamaño $n_I = 2$ mediante un diseño de muestreo sin reemplazo tal que

$$p_I(s_I) = \begin{cases} 0.5, & \text{si } s_I=\{U_1, U_2\},\\ 0.4, & \text{si } s_I=\{U_1, U_3\},\\ 0.1, & \text{si } s_I=\{U_2, U_3\} \end{cases}$$

Ahora, suponga que dentro de cada unidad primaria seleccionada se selecciona un solo elemento de acuerdo a los siguientes diseños de muestreo

$$p_1(S_1 \mid S_I) = \begin{cases} 0.5, & \text{si } s_1 = \{Yves\}, \\ 0.5, & \text{si } s_1 = \{Ken\} \end{cases}$$

$$p_2(S_2 \mid S_I) = \begin{cases} 0.9, & \text{si } s_2 = \{Erik\}, \\ 0.1, & \text{si } s_2 = \{Sharon\} \end{cases}$$

$$p_3(S_3 \mid S_I) = \begin{cases} 1.0, & \text{si } s_3 = \{Leslie\} \end{cases}$$

Es decir, el tamaño de la muestra final es $n = 2$. Y el soporte de la primera etapa está dado por

$$Q_I = \{\{U_1, U_2\}, \{U_1, U_3\}, \{U_2, U_3\}\},$$

y los soportes de la segunda etapa están dados por $Q^1 = \{\{\text{Yves}\}, \{\text{Ken}\}\}$, $Q^2 = \{\{\text{Erick}\}, \{\text{Sharon}\}\}$ y $Q^3 = \{\{\text{Leslie}\}\}$. Dado lo anterior, el soporte Q está dada por

$$Q = \left\{ \bigcup_{i \in s_I^{(1)}} s_i, \bigcup_{i \in s_I^{(2)}} s_i, \bigcup_{i \in s_I^{(3)}} s_i \right\},$$

donde

$$\bigcup_{i \in s_I^{(1)}} s_i = \{\{\text{Yves}, \text{Erick}\}, \{\text{Yves}, \text{Sharon}\}, \{\text{Ken}, \text{Erick}\}, \{\text{Ken}, \text{Sharon}\}\},$$

$$\bigcup_{i \in s_I^{(2)}} s_i = \{\{\text{Erick}, \text{Leslie}\}, \{\text{Sharon}, \text{Leslie}\}\},$$

y

$$\bigcup_{i \in s_I^{(3)}} s_i = \{\{\text{Yves}, \text{Leslie}\}, \{\text{Ken}, \text{Leslie}\}\}.$$

Las probabilidades $\prod_{i \in s_I} p_i(s_i)$ y $p_I(s_I)$ para todas las posibles muestras son como sigue a continuación:

```
                   p(s_1) X p(s_2)        p(s_I)          p(s)
Yves    Erick           0.5 X 0.9            0.5          0.225
Yves    Sharon          0.5 X 0.1            0.5          0.025
Ken     Erick           0.5 X 0.9            0.5          0.225
Ken     Sharon          0.5 X 0.1            0.5          0.025
Erick   Leslie          0.9 X 1.0            0.1          0.090
Sharon  Leslie          0.1 X 1.0            0.1          0.010
Yves    Leslie          0.5 X 1.0            0.4          0.200
Ken     Leslie          0.5 X 1.0            0.4          0.200

Total                                                     1.000
```

Se observa que $p(s)$ es un auténtico diseño de muestreo. Nótese que dentro de cada posible muestra de la primera etapa, la suma de probabilidades es igual a uno. Por ejemplo, para $S_I = \{U_1, U_2\}$, las posibles muestras en la segunda etapa corresponden a {Yves, Erick}, {Yves, Sharon}, {Ken, Erick} y {Ken, Sharon} con probabilidades 0.45, 0.05, 0.45 y 0.05, respectivamente, y la suma de estas probabilidades es igual a uno.

Los parámetros poblacionales de interés pueden escribirse como:

1. El total poblacional,

$$t_y = \sum_{k\in U} y_k = \sum_{i=1}^{N_I}\sum_{k\in U_i} y_k = \sum_{i=1}^{N_I} t_{yi} \tag{7.1.6}$$

donde $t_{yi} = \sum_{k\in U_i} y_k$ es el total de la i-ésima unidad primaria de muestreo $i = 1, \ldots, N_I$.

2. La media poblacional,

$$\bar{y}_U = \frac{\sum_{k\in U} y_k}{N} = \frac{1}{N}\sum_{i=1}^{N_I}\sum_{k\in U_i} y_k = \frac{1}{N}\sum_{i=1}^{N_I} N_i\bar{y}_i \tag{7.1.7}$$

donde $\bar{y}_i = \dfrac{1}{N_i}\sum_{k\in U_i} y_k$ es la media de la i-ésima unidad primaria de muestreo $i = 1, \ldots, N_I$.

Ejemplo 7.1.2. Nuestra población ejemplo U_I dada por

$$U_I = \{U_1, U_2, U_3\}$$

Suponga que se selecciona una muestra s_I de unidades primarias de muestreos de tamaño $n_I = 2$. El sub-muestreo en la segunda etapa es tal que en cada unidad primaria de muestreo seleccionada en la primera etapa se selecciona un sólo elemento, de tal forma que el tamaño de la muestra de elementos es de dos. Defina el soporte Q de elementos si la selección de la muestra es con reemplazo.

7.1.1 El estimador de Horvitz-Thompson

En la primera etapa las probabilidades de inclusión de primer y segundo orden, de las unidades primarias de muestreo, inducidas por el diseño de muestreo $p_I(s_I)$ están dadas por π_{Ii} y π_{Iij} respectivamente con $i, j \in U_I$. Por tanto se tiene que

$$\Delta_{Iij} = \begin{cases} \pi_{Iij} - \pi_{Ii}\pi_{Ij}, & \text{si } i, j \in U_I, \\ \pi_{Ii}(1 - \pi_{Ii}), & \text{si } i = j \in U_I. \end{cases} \tag{7.1.8}$$

En la segunda etapa las probabilidades de inclusión de primer y segundo orden, de los elementos en la i-ésima $i \in S_I$ unidad primaria de muestreo, inducidas por el diseño de muestreo $p_i(s_i)$ y condicionadas a que U_i fue seleccionada en la muestra de la primera etapa están dadas por $\pi_{k|i}$ y $\pi_{kl|i}$ respectivamente para $k, l \in U_i$ con

$\pi_{k|i} = Pr(k \in S_i|U_i \in S_I)$ y $\pi_{kl|i} = Pr(k \in S_i, l \in S_i|U_i \in S_I)$. Por tanto se tiene que

$$\Delta_{kl|i} = \begin{cases} \pi_{kl|i} - \pi_{k|i}\pi_{l|i}, & \text{si } k \neq l, \\ \pi_{k|i}(1 - \pi_{k|i}), & \text{si } k = l. \end{cases} \tag{7.1.9}$$

En general, de la definición de probabilidad de inclusión se tiene el siguiente resultado.

Resultado 7.1.2. *La probabilidad de inclusión de primer orden del k-ésimo elemento de U está dada por*

$$\begin{aligned} \pi_k = Pr(k \in S) &= Pr(k \in S_i \ y \ i \in S_I) \\ &= Pr(k \in S_i|i \in S_I)Pr(i \in S_I) = \pi_{k|i}\pi_{Ii} \end{aligned} \tag{7.1.10}$$

La probabilidad de inclusión de segundo orden está dada por

$$\pi_{kl} = \begin{cases} \pi_{Ii}\pi_{k|i}, & \text{si } k = l \in U_i, \\ \pi_{Ii}\pi_{k|i}, & \text{si } k \neq l \in U_i, \\ \pi_{Iij}\pi_{k|i}\pi_{l|j}, & \text{si } k \in U, l \in U_j (i \neq l). \end{cases} \tag{7.1.11}$$

Con el anterior resultado podemos utilizar la forma general del estimador de Horvitz-Thompson para hallar su expresión particular y su varianza bajo un diseño de muestreo bietápico (Särndal, Swensson & Wretman 1992). Sin embargo, para hallar una forma más rápida de calcular la varianza del estimador necesitamos recurrir a algunos resultados muy conocidos de la teoría de probabilidad. Éstos han sido utilizados ampliamente en el campo del muestreo, pero no fue sino hasta que Hansen, Hurwitz & Madow (1953) publicaron dichos resultados aplicados al muestreo. En general, se trata de expresar:

- La esperanza de una variable aleatoria como el valor esperado de esperanzas condicionales.
- La varianza de una variable aleatoria como la suma de la varianza de esperanzas condicionales y la esperanza de varianzas condicionales.

Resultado 7.1.3. *Sean U y H variables aleatorias, entonces:*

$$E_1(U) = E_2(E_1(U|H)) \tag{7.1.12}$$

y, a su vez,

$$Var_1(U) = E_2(Var_1(U|H)) + Var_2(E_1(U|H)) \tag{7.1.13}$$

En donde el subíndice 1, denota la esperanza o varianza inducida por la función de distribución de la variable aleatoria U, y el subíndice 2 denota la esperanza o varianza inducida por la función de distribución de la variable aleatoria H.

Prueba. Es necesario recordar que $Pr(U = U_i|H_j) = Pr(U = U_i, H = H_j)/Pr(H_j)$ y además que $Pr(U = U_i) = \sum_j (U = U_i, H = H_j)$, por consiguiente.

1. Esperanza:

$$\begin{aligned}
E_1(U) &= \sum_i U_i Pr(U = U_i) \\
&= \sum_i U_i \sum_j Pr(U = U_i, H = H_j) \\
&= \sum_i U_i \sum_j Pr(U = U_i | H = H_j) Pr(H = H_j) \\
&= \sum_j Pr(H = H_j) \sum_i U_i Pr(U = U_i | H = H_j) \\
&= \sum_j Pr(H = H_j) E_2(U | H = H_j) \\
&= E_2(E_1(U|H))
\end{aligned}$$

2. Covarianza: sea W, una variable aleatoria y tomemos a $x = E_2(U)$ y $y = E_2(W)$

$$\begin{aligned}
Cov(U, W) &= E(UW) - E(U)E(W) \\
&= E_1(E_2(UW)) - E_1(E_2(U))E_1(E_2(W)) \\
&= E_1(E_2(UW)) - E_1(x)E_1(y) \\
&= E_1\left[E_2(UW) - xy\right] + E(xy) - E_1(x)E_1(y) \\
&= E_1\left[Cov_2(U, W)\right] + Cov_1(x, y) \\
&= E_1\left[Cov_2(U, W)\right] + Cov_1\left[E_2(U), E_2(W)\right]
\end{aligned}$$

3. Varianza: dado que la varianza es un caso particular de la covarianza, entonces:

$$\begin{aligned}
Var(U) = Cov(U, U) &= E_1\left[Cov_2(U, U)\right] + Cov_1\left[E_2(U), E_2(U)\right] \\
&= E_1[Var_2(U)] + Var_1[E_2(U)]
\end{aligned}$$

■

Con ayuda del anterior resultado es posible obtener expresiones para el estimador de Horvitz-Thompson que muestren la variación en cada una de las dos etapas de este diseño de muestreo. Es interesante la forma que toma tanto el estimador genérico como su respectiva varianza porque, dado que existen dos etapas de muestreo, en la primera se estiman los totales de los conglomerados y, en la segunda etapa se estima el gran total utilizando esas estimaciones en las unidades primarias seleccionadas. Como el proceso de estimación se lleva a cabo en dos etapas, es de esperarse que existan dos fuentes de variación: la primera debido a la estimación de los totales de las unidades primarias de muestreo y la segunda debido a la estimación del gran total. Suponiendo que fueron seleccionadas cuatro unidades primarias de muestreo, existirán entonces cuatro estimaciones cuya varianza estará sintetizada en una sola expresión, mientras que, por otro lado, existirá otra fuente de variación cuando se quiera estimar el gran total.

Resultado 7.1.4. *Bajo muestreo en dos etapas el estimador de Horvitz-Thompson es insesgado para el total poblacional y toma la forma*

$$\hat{t}_{y,\pi} = \sum_{i \in S_I} \sum_{k \in S_i} \frac{y_k}{\pi_{Ii}\pi_{k|i}} = \sum_{i \in S_I} \frac{\hat{t}_{yi,\pi}}{\pi_{Ii}} \tag{7.1.14}$$

con varianza dada por

$$Var_{BI}(\hat{t}_{y,\pi}) = \underbrace{\sum\sum_{U_I} \Delta_{Iij} \frac{t_i}{\pi_{Ii}} \frac{t_j}{\pi_{Ij}}}_{Var(UPM)} + \underbrace{\sum_{i \in U_I} \frac{Var_{p_i}(\hat{t}_i)}{\pi_{Ii}}}_{Var(USM)} \tag{7.1.15}$$

cuya estimación insesgada es

$$\widehat{Var}_{BI}(\hat{t}_{y,\pi}) = \underbrace{\sum\sum_{S_I} \frac{\Delta_{Iij}}{\pi_{Iij}} \frac{\hat{t}_{yi,\pi}}{\pi_{Ii}} \frac{\hat{t}_{yj,\pi}}{\pi_{Ij}}}_{\widehat{Var}(UPM)} + \underbrace{\sum_{i \in S_I} \frac{\widehat{Var}(\hat{t}_{yi,\pi})}{\pi_{Ii}}}_{\widehat{Var}(USM)} \tag{7.1.16}$$

donde

$$Var(\hat{t}_i) = \sum\sum_{U_i} \Delta_{kl|i} \frac{y_k}{\pi_{k|i}} \frac{y_l}{\pi_{l|i}} \tag{7.1.17}$$

$$\hat{t}_{yi,\pi} = \sum_{k \in S_i} \frac{y_k}{\pi_{k|i}}$$

representando la estimación del total de la característica de interés en la i-ésima unidad primaria de muestreo y

$$\widehat{Var}(\hat{t}_i) = \sum\sum_{S_i} \frac{\Delta_{kl|i}}{\pi_{kl|i}} \frac{y_k}{\pi_{k|i}} \frac{y_l}{\pi_{l|i}} \tag{7.1.18}$$

Nótese que la variación del estimador se descompone en las dos etapas propias de este diseño. Además es importante tener en cuenta que $\widehat{Var}(UPM)$ y $\widehat{Var}(USM)$ no son estimadores insesgados para $Var(UPM)$ y $Var(USM)$ respectivamente. Sin embargo, toda la expresión $\widehat{Var}_{BI}(\hat{t}_{y,\pi})$ sí lo es para $Var_{BI}(\hat{t}_{y,\pi})$.

Prueba. Para desarrollar el anterior resultado es necesario manejar los dos conceptos inherentes al muestreo en dos o más etapas. **a)La invarianza:** para seleccionar las unidades primarias de muestreo se debe utilizar un mismo diseño y **b)La independencia:** cualquiera que fuere el diseño escogido para seleccionar los elementos dentro de una unidad primaria de muestreo, éste no debe afectar el sub-muestreo en cualquier otra unidad primaria de muestreo; por tanto, cualquier covarianza existente en esta etapa será nula.

En primer lugar, se tiene la siguiente forma para el estimador de Horvitz-Thompson:

$$\begin{aligned}
\hat{t}_{y,\pi} &= \sum_{k\in S}\frac{y_k}{\pi_k} && (7.1.19)\\
&= \sum_{i\in S_I}\sum_{k\in S_i}\frac{y_k}{\pi_{Ii}\pi_{k|i}} && (7.1.20)\\
&= \sum_{i\in S_I}\frac{1}{\pi_{Ii}}\sum_{k\in S_i}\frac{y_k}{\pi_{k|i}} && (7.1.21)\\
&= \sum_{i\in S_I}\frac{\hat{t}_{yi,\pi}}{\pi_{Ii}} && (7.1.22)
\end{aligned}$$

1. Insesgamiento del estimador:

$$\begin{aligned}
E_p(\hat{t}_{y,\pi}) &= E_{p_I}\left(E_p\left[\sum_{i\in S_I}\frac{\hat{t}_{yi,\pi}}{\pi_{Ii}}\mid S_I\right]\right)\\
&= E_{p_I}\left(\sum_{i\in S_I}\underbrace{E_p\left[\frac{\hat{t}_{yi,\pi}}{\pi_{Ii}}\mid S_I\right]}_{\text{invarianza}}\right)\\
&= E_{p_I}\left(\sum_{i\in S_I}\frac{E_{p_i}(\hat{t}_{yi,\pi})}{\pi_{Ii}}\right)\\
&= E_{p_I}\left(\sum_{i\in S_I}\frac{t_{yi,\pi}}{\pi_{Ii}}\right)\\
&= \sum_{i\in U_I}\frac{t_{yi,\pi}}{\pi_{Ii}}E_{P_I}(I_{Ii}(S_I)) = t_y
\end{aligned}$$

2. Varianza:

$$Var_p(\hat{t}_{y,\pi}) = \underbrace{Var_{p_I}\left(E_p\left[\hat{t}_{y,\pi}\mid S_I\right]\right)}_{Var(UPM)} + \underbrace{E_{p_I}\left(Var_p\left[\hat{t}_{y,\pi}\mid S_I\right]\right)}_{Var(USM)} \qquad (7.1.23)$$

El primer sumando es equivalente a

$$
\begin{aligned}
Var_{p_I}\left(E_p\left[\hat{t}_{y,\pi} \mid S_I\right]\right) &= Var_{p_I}\left(E_p\left[\sum_{i\in S_I}\frac{\hat{t}_{yi,\pi}}{\pi_{Ii}} \mid S_I\right]\right)\\
&= Var_{p_I}\left(\sum_{i\in S_I}\underbrace{\frac{E_p(\hat{t}_{y,\pi} \mid S_I)}{\pi_{Ii}}}_{Invarianza}\right)\\
&= Var_{p_I}\left(\sum_{i\in S_I}\frac{E_p(\hat{t}_{y,\pi})}{\pi_{Ii}}\right)\\
&= Var_{p_I}\left(\sum_{i\in S_I}\frac{t_{yi,\pi}}{\pi_{Ii}}\right)\\
&= \sum\sum_{U_I}\Delta_{Iij}\frac{t_{yi,\pi}}{\pi_{Ii}}\frac{t_{yj,\pi}}{\pi_{Ij}}
\end{aligned}
$$

El segundo sumando toma la siguiente forma

$$
\begin{aligned}
E_{p_I}\left(Var_p\left[\hat{t}_{y,\pi} \mid S_I\right]\right) &= E_{p_I}\left(Var_p\left[\sum_{i\in S_I}\frac{\hat{t}_{yi,\pi}}{\pi_{Ii}} \mid S_I\right]\right)\\
&= E_{p_I}\left(\sum_{i\in S_I}\frac{Var_p(\hat{t}_{yi,\pi} \mid S_I)}{\pi^2_{Ii}}\right)\\
&= E\left(\sum_{i\in S_I}\left[\frac{Var(\hat{t}_{yi,\pi})}{\pi^2_{Ii}}\right]\right)\\
&= E_{p_I}\sum_{i\in U_I}\frac{I_{Ii}(S_I)}{\pi^2_{Ii}}Var_{p_i}(\hat{t}_{yi,\pi})\\
&= \sum_{i\in U_I}\left[\frac{Var(\hat{t}_{yi,\pi})}{\pi_{Ii}}\right]
\end{aligned}
$$

Luego, la varianza del estimador está dada por la expresión (7.1.15).

3. Varianza Estimada: para verificar que $\widehat{Var}_{BI}(\hat{t}_{y,\pi})$ es un estimador insesgado de la varianza del estimador de Horvitz-Thompson, se debe tener en cuenta que

$$
\begin{aligned}
E\left(\hat{t}_{yi,\pi}\hat{t}_{yj,\pi} \mid S_I\right) &= \begin{cases} Var_{p_i}(\hat{y}_{yi,\pi}) + (E_{p_i}(\hat{y}_{yi,\pi}))^2, & \text{si } i=j,\\ E_{p_i}(\hat{y}_{yi,\pi})E_{p_j}(\hat{y}_{yj,\pi}), & \text{si } i\neq j\end{cases}\\
&= \begin{cases} Var(\hat{t}_{yi,\pi}) + t^2_{yi,\pi}, & \text{si } i=j,\\ (t_{yi,\pi})(t_{yj,\pi}), & \text{si } i\neq j\end{cases}
\end{aligned} \tag{7.1.24}
$$

Para la primera parte de la varianza estimada se tiene que

$$
\begin{aligned}
&E_{p_I}\left(E_p\left[\sum\sum_{S_I}\frac{\Delta_{Iij}}{\pi_{Iij}}\frac{\hat{t}_{yi,\pi}}{\pi_{Ii}}\frac{\hat{t}_{yj,\pi}}{\pi_{Ij}} \mid S_I\right]\right)\\
&=E_{p_I}\sum\sum_{S_I}\frac{\Delta_{Iij}}{\pi_{Iij}}\frac{E_p(\hat{t}_{yi,\pi}\hat{t}_{yj,\pi}\mid S_I)}{\pi_{Ii}\pi_{Ij}}\\
&=E\left(\sum_{i\in S_I}\sum_{j\neq i\in S_I}\frac{\Delta_{Iij}}{\pi_{Iij}}\frac{(t_{yi,\pi})}{\pi_{Ii}}\frac{(t_{yj,\pi})}{\pi_{Ij}}+\sum_{S_I}\frac{\Delta_{Iii}}{\pi_{Iii}}\frac{Var(\hat{t}_{yi,\pi})+t^2_{yi,\pi}}{\pi^2_{Ii}}\right)\\
&=E\left(\sum_{i\in S_I}\sum_{j\in S_I}\frac{\Delta_{Iij}}{\pi_{Iij}}\frac{(t_{yi,\pi})}{\pi_{Ii}}\frac{(t_{yj,\pi})}{\pi_{Ij}}+\sum_{S_I}\frac{Var(\hat{t}_{yi,\pi})}{\pi^2_{Ii}}(1-\pi_{Ii})\right)\\
&=\sum_{i\in U_I}\sum_{j\in U_I}\Delta_{Iij}\frac{(t_{yi,\pi})}{\pi_{Ii}}\frac{(t_{yj,\pi})}{\pi_{Ij}}-\sum_{U_I}Var(\hat{t}_{yi,\pi})\left(1-\frac{1}{\pi_{Ii}}\right)
\end{aligned}
$$

Para la segunda parte de la varianza estimada se tiene que

$$
\begin{aligned}
&E\left(E\left[\sum_{i\in S_I}\frac{\widehat{Var}(\hat{t}_{yi,\pi})}{\pi_{Ii}}\mid S_I\right]\right)\\
&=E\left(\sum_{i\in S_I}\frac{Var(\hat{t}_{yi,\pi})}{\pi_{Ii}}\right)\\
&=\sum_{i\in U_I}Var(\hat{t}_{yi,\pi})\\
&=\sum_{U_I}\frac{Var(\hat{t}_{yi,\pi})}{\pi_{Ii}}+\sum_{U_I}Var(\hat{t}_{yi,\pi})\left(1-\frac{1}{\pi_{Ii}}\right)
\end{aligned}
$$

Sumando estas dos cantidades se llega al resultado. Nótese que por sí solas, estas cantidades no son insesgadas para sus contrapartes poblacionales, sin embargo se tiene que:

$$
E\left[\widehat{Var}(UPM)\right]+E\left[\widehat{Var}(USM)\right]=Var(\hat{t}_{y,\pi}) \tag{7.1.25}
$$

■

Al respecto de la forma que toma la varianza del estimador de Horvitz-Thompson, Särndal, Swensson & Wretman (1992) afirman que:

- Es conveniente estimar los dos componentes de varianza $Var(UPM)$ y $Var(USM)$ separadamente para tener una idea del aporte de variabilidad en cada una de las etapas.
- Si $\pi_{k|i}=\pi_{kl|i}=1$ para todo $k,l\in U_i$ y para todo $U_i\in S_I$, entonces $Var(USM)=0$ entonces este diseño toma la forma de un diseño de conglomerados.

- Si $\pi_{Ii} = \pi_{Iij} = 1$ para todo $i, j = 1, \ldots, N_I$, entonces este diseño se torna en un diseño estratificado.

Ejemplo 7.1.3. Utilizando la información del ejemplo 7.1.1, compruebe, mediante un ejercicio léxico-gráfico, el insesgamiento del estimador de Horvitz-Thompson.

7.2 Diseño de muestreo MAS-MAS

En el muestreo aleatorio simple de conglomerados se medían todos y cada una de los elementos pertenecientes a los conglomerados seleccionados en la muestra s_I. Sin embargo, dado que, en la mayoría de situaciones, los conglomerados tienden a ser muy similares en el comportamiento estructural de la característica de interés se consideraría un desperdicio de recursos económicos y logísticos la incorporación de elementos que no traen consigo nueva información. Para esto es más económico tomar una muestra más amplia de unidades primarias de muestro y realizar un sub-muestreo dentro de cada una de ellas.

Este diseño de muestreo supone que la población está divida en N_I unidades primarias de muestreo, de las cuales se selecciona una muestra s_I de n_I unidades mediante un diseño de muestreo aleatorio simple. El sub-muestreo dentro de cada unidad primaria seleccionada es también aleatorio simple. Es decir, para cada unidad primaria de muestreo seleccionada $i \in s_{Ih}$ de tamaño N_i se selecciona una muestra s_i de elementos de tamaño n_i.

7.2.1 Algoritmos de selección

En la selección de las muestras de unidades primarias y secundarias sin reemplazo se utilizan los algoritmos de muestreo dados en el capítulo 2, de tal forma que los siguientes pasos se deben realizar:

- Separar la población en N_I unidades primarias de muestreo mediante el marco de muestreo de conglomerados.
- Realizar una selección de n_I conglomerados mediante cualquiera de los métodos expuestos en la sección 3.2.1; es decir, por el método coordinado negativo o por el método de Fan-Muller-Rezucha.
- Para cada unidad primaria seleccionada en la muestra de la primera etapa s_I, realizar una selección de n_i $i \in S_I$ elementos mediante cualquiera de los métodos expuestos en la sección 3.2.1.

Resultado 7.2.1. *Cuando el diseño de muestreo es aleatorio simple en las dos etapas, se tienen las siguientes probabilidades de inclusión de primer y segundo orden*

$$\pi_{Ii} = \frac{n_I}{N_I} \tag{7.2.1}$$

$$\pi_{Iij} = \frac{n_I(n_I-1)}{N_I(N_I-1)} \tag{7.2.2}$$

respectivamente. Por otro lado, la probabilidad de inclusión de un elemento o unidad secundaria de muestreo perteneciente a la i-ésima unidad primaria de muestreo $i \in U_I$ está dado por

$$\pi_k = \frac{n_I}{N_I}\frac{n_i}{N_i} \tag{7.2.3}$$

Una vez que la muestra de unidades primarias s_I es seleccionada se dispone a realizar una enumeración completa de los elementos pertenecientes a ésta para levantar un marco de muestreo que permita la selección de una sub-muestra para realizar la respectiva medición de todos y cada uno de los elementos pertenecientes a la sub-muestra seleccionada. En el diseño de muestreo aleatorio por conglomerados el estimador del total poblacional t_y estaba dado por $\hat{t}_{y,\pi} = \frac{N_i}{n_i}\sum_{i\in S_I} t_{yi}$ porque se conocían los totales exactos de cada conglomerado seleccionado mediante la realización de un censo en los mismos. Por otra parte, en el muestreo en dos etapas MAS-MAS, debido a que no se miden todos los elementos de las unidades primarias seleccionadas, se deben estimar estos totales t_{yi} mediante la siguiente expresión

$$\hat{t}_{yi,\pi} = \frac{N_i}{n_i}\sum_{k\in S_i} y_k = N_i\bar{y}_{U_i} \tag{7.2.4}$$

Con el siguiente resultado se llega a una estimación del parámetro de interés

Resultado 7.2.2. *Bajo muestreo en dos etapas MAS-MAS, el estimador de Horvitz-Thompson es insesgado para el total poblacional y toma la forma*

$$\hat{t}_{y,\pi} = \frac{N_I}{n_I}\sum_{i\in S_I}\frac{N_i}{n_i}\sum_{k\in S_i} y_k \tag{7.2.5}$$

con varianza dada por

$$Var_{MM}(\hat{t}_{y,\pi}) = \frac{N_I^2}{n_I}\left(1-\frac{n_I}{N_I}\right)S^2_{t_yU_I} + \frac{N_I}{n_I}\sum_{i\in U_I}\frac{N_i^2}{n_i}\left(1-\frac{n_i}{N_i}\right)S^2_{y_{U_i}} \tag{7.2.6}$$

cuya estimación insesgada es

$$\widehat{Var}_{MM}(\hat{t}_{y,\pi}) = \frac{N_I^2}{n_I}\left(1-\frac{n_I}{N_I}\right)S^2_{\hat{t}_yS_I} + \frac{N_I}{n_I}\sum_{i\in S_I}\frac{N_i^2}{n_i}\left(1-\frac{n_i}{N_i}\right)S^2_{y_{S_i}} \tag{7.2.7}$$

donde $S^2_{t_yU_I}$ es la varianza poblacional de los totales t_{yi} $i \in U_I$ de todas y cada una de las unidades primarias de muestreo y $S^2_{y_{U_i}}$ es la varianza poblacional entre los elementos dentro de cada unidad primaria de muestreo. Similarmente, $S^2_{\hat{t}_y s_I}$ y $S^2_{y_{s_i}}$.

El primer término de (7.2.6) se refiere a la variabilidad debida a la primera etapa del diseño muestral mientras que el segundo sumando se refiere a la varianza adicional debida al sub-muestreo en las unidades primarias de muestreo. Lohr (2000) afirma que, de igual manera como en el caso del diseño de muestreo por

conglomerados, si las unidades primarias de muestreo presentan distintos tamaños entonces la variabilidad del estimador puede ser muy grande. Si los tamaños N_i de los conglomerados $i \in U_I$ son muy diferentes entre sí, el componente de varianza será grande incluso si el comportamiento estructural de la característica de interés es constante en cada unidad primaria.

7.2.2 Tamaño de muestra

Cada vez que avanzamos en el desarrollo programático de este texto nos encontramos, si bien los principios de estimación son los mismos, con que el diseño de la encuesta y la estimación de los parámetros de interés se tornan más complejos. Lohr (2000) afirma que la mejor manera de diseñar una encuesta es revisarla después de que esta haya concluido pues, al finalizar la encuesta, es posible evaluar el efecto de las unidades primarias de muestreo sobre la estimación final y, de esta manera, es posible saber en dónde se deberían asignar más recursos logísticos para obtener una mejor información. Pero a pesar de que el conocimiento de la población sea aceptable, siempre surge la pregunta del tamaño de muestra. En particular, ¿cuántas unidades primarias de muestreo se deberían seleccionar en la muestra? y ¿cuántos elementos o unidades secundarias de muestreo deberían ser seleccionados en el sub-muestreo dentro de las unidades primarias de muestreo?

Por ejemplo, en particular en las encuestas de áreas mientras mayor sea el tamaño de la unidad primaria de muestreo, se puede esperar que exista más variabilidad de dentro de la misma. Sin embargo, si el tamaño de unidad primaria es muy grande, se podrían perder los beneficios del ahorro financiero y logístico.

El objetivo de una buena muestra es recopilar la mayor cantidad de información al menor precio económico y operativo. Suponga que la población está divida en N_I unidades primarias de muestreo, de las cuales se selecciona una muestra s_I de n_I unidades. Cada unidad primaria de muestreo contiene exactamente $N_i = M$ elementos o unidades secundarias de muestreo. El sub-muestreo es tal que se selecciona una muestra de exactamente $n_i = m$ unidades secundarias de muestreo. Por tanto, el tamaño poblacional y muestral estará dado por

$$N = N_I M \qquad \text{y} \qquad n = n_I m \tag{7.2.8}$$

respectivamente. De tal forma que el estimador de t_y se puede escribir como

$$\hat{t}_{y,\pi} = \frac{N_I}{n_I}\frac{M}{m}\sum_{i\in S_I}\sum_{k\in S_i} y_k \tag{7.2.9}$$

y su varianza como

$$Var_{MM}(\hat{t}_{y,\pi}) = \frac{N_I^2}{n_I}\left(1-\frac{n_I}{N_I}\right)S^2_{t_yU_I} + \frac{N_I^2M^2}{n_Im}\left(1-\frac{m}{M}\right)\bar{S}^2_{yU_i} \tag{7.2.10}$$

donde $\bar{S}^2_{yU_i} = (1/N_I)\sum_{i\in U_I} S^2_{yU_i}$.

Resultado 7.2.3. *Utilizando los resultados de la descomposición de las sumas de cuadrados, la varianza de la estrategia en dos etapas (2MAS) toma la siguiente forma*

$$Var_{2MAS}(\hat{t}_{y,\pi}) = \frac{N_I^2 M}{n_I}\left[\frac{1}{N_I-1}(SCT - SCD) + \left(\frac{M}{m}-1\right)\frac{SCD}{N_I(M-1)}\right] \tag{7.2.11}$$

mientras que la varianza de la estrategia aleatoria simple, con un tamaño poblacional igual a $N = M \times N_I$ elementos y un tamaño de muestra igual a $n = m \times n_I$ elementos, se puede escribir como

$$Var_{MAS}(\hat{t}_{y,\pi}) = \frac{N_I^2}{n_I}\left(1-\frac{n_I}{N_I}\right)M\frac{SCT}{MN_I-1} \tag{7.2.12}$$

Para encontrar los valores óptimos de n_I y m que serán utilizados en la primera y segunda etapa de muestro de tal forma que dada una función de costo se minimice[2] la varianza del estimador. Por tanto, se tiene el siguiente resultado.

Resultado 7.2.4. *Al considerar la siguiente función de costo*

$$C = c_1 n_I + c_2 n_I m \tag{7.2.13}$$

donde c_1 es el costo de del levantamiento del marco de muestreo en cada unidad primaria seleccionada en la muestra s_I y c_2 es el costo de recolectar la información de la característica de interés para los elementos o unidades secundarias seleccionadas por el sub-muestreo. Los valores óptimos de n_I y m que minimizan la varianza del estimador dada por la expresión (7.2.6) restringido al costo total de la encuesta dado por (7.2.11) son

$$n_I = \frac{C}{c_1 + c_2 m} \tag{7.2.14}$$

y

$$m = M\bar{S}^2_{yU_i}\sqrt{\frac{c_1/c_2}{S^2_{t_yU_I} - M\bar{S}^2_{yU_i}}} \tag{7.2.15}$$

Prueba. La cantidad a minimizar está dada en la expresión (7.2.10) que está sujeta a la restricción de la función de costo (7.2.11). Utilizando el método de los multiplicadores de Lagrange, se tiene que

$$\mathcal{L}(n_I, m, \lambda) = \frac{N_I^2}{n_I}\left(1-\frac{n_I}{N_I}\right)S^2_{t_yU_I} + \frac{N_I^2M^2}{n_I m}\left(1-\frac{m}{M}\right)\bar{S}^2_{yU_i} + \lambda(c_1 n_I + c_2 n_I m - C) \tag{7.2.16}$$

Anulando las derivadas parciales se tiene que

$$\frac{\partial\mathcal{L}}{\partial n_I} = -\frac{N_I^2M^2}{n_I^2}\left(\frac{1}{m}-\frac{1}{M}\right)\bar{S}^2_{yU_i} - \frac{N_I^2}{n_I^2}S^2_{t_yU_I} + c_1\lambda + c_2 m\lambda = 0 \tag{7.2.17}$$

$$\frac{\partial\mathcal{L}}{\partial m} = -\frac{N_I^2M^2}{n_I^2m^2}\bar{S}^2_{yU_i} + c_2 n_I\lambda = 0 \tag{7.2.18}$$

[2]Naturalmente estos valores dependerán de la función de costo utilizada.

De (7.2.15) se tiene que

$$n_I^2 = -\frac{N_I^2 M^2 \left(\frac{1}{m} - \frac{1}{M}\right) \bar{S}^2_{y_{U_i}} + N_I^2 S^2_{t_y U_I}}{c_1\lambda + c_2 m\lambda} \tag{7.2.19}$$

De (7.2.16) se tiene que

$$n_I^2 = -\frac{N_I^2 M^2 \bar{S}^2_{y_{U_i}}}{c_2 m^2 \lambda} \tag{7.2.20}$$

Igualando las anteriores ecuaciones y despejando m se tiene la demostración del resultado. ■

Si $\bar{S}^2_{y_{U_i}}$, la variabilidad de la característica de interés dentro de las unidades primarias es grande, entonces m será grande. Se debe resaltar que los resultados son válidos si la función de costo es la correcta.

7.2.3 Estimación de la varianza en muestreo de dos etapas

Cuando la estrategia de muestreo hace uso del estimador de Horvitz-Thompson podemos utilizar su forma general para hallar su varianza bajo cualquier diseño de muestreo. La expresión de la varianza del estimador de Horvitz-Thompson bajo muestreo bietápico está dada por

$$Var(\hat{t}_\pi) = \sum\sum_{UI} \Delta_{Iij} \frac{t_j}{\pi_{Ij}} \frac{t_i}{\pi_{Ii}} + \sum_{UI} V_i/\pi_{Ii} \tag{7.2.21}$$

cuya estimación insesgada es

$$\widehat{Var}_1(\hat{t}_\pi) = \sum\sum_{sI} \frac{\Delta_{Iij}}{\pi_{Ii}} \frac{\hat{t}_i}{\pi_{Ii}} \frac{\hat{t}_j}{\pi_{Ij}} + \sum_{sI} \hat{V}_i/\pi_{Ii} \tag{7.2.22}$$

La expresión anterior involucra el cálculo de las varianzas de las variables dentro de cada conglomerado. Lo anterior en una encuesta a gran escala puede llegar a ser muy tedioso, costoso y además muy demorado. Särndal, Swensson & Wretman (1992, p. 139) dan una posible solución al problema, ésta es mantener la primera parte del estimador de la varianza como estimador general de la misma. Así, un estimador sencillo, pero sesgado, es

$$\widehat{Var}_2(\hat{t}_\pi) = \sum\sum_{sI} \frac{\Delta_{Iij}}{\pi_{Ii}} \frac{\hat{t}_i}{\pi_{Ii}} \frac{\hat{t}_j}{\pi_{Ij}} \tag{7.2.23}$$

El anterior estimador sobre-estima la varianza para las unidades primarias de muestreo, pero a su vez también lo hace con (7.2.19). Otra posible solución para estimar la varianza del estimador de Horvitz-Thompson, es asumir que el muestreo

en la primera etapa se llevó a cabo con reemplazo. Así, la estimación (sesgada) de la varianza estaría dada por

$$\widehat{Var}_3(\hat{t}_\pi) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\frac{\hat{t}_i}{p_{Ii}} - \hat{t}_\pi \right)^2 \tag{7.2.24}$$

Un caso especial del anterior término, se tiene suponiendo que $\pi_k = np_k$, y si el muestreo en la primera etapa fue aleatorio simple, entonces $p_k = \frac{1}{N}$. El estimador de la varianza, bajo la anterior condición es

$$\widehat{Var}(\hat{t}_\pi) = \frac{N^2}{n(n-1)} \sum_{i=1}^{n} \left(\hat{t}_i - \frac{\sum_{i=1}^{n} \hat{t}_i}{n} \right)^2 = \frac{N^2}{n} S^2_{\hat{t}_i}$$

Srinath & Hidiroglou (1980) proponen un método *rápido* para la estimación de la varianza del estimador de Horvitz-Thompson. Éste supone que el método de selección en la segunda etapa es MAS y es invariante en la primera etapa (se puede seleccionar la muestra en la primera etapa mediante cualquier diseño); lo que conlleva a que este estimador de la varianza sea insesgado y está dado por

$$\widehat{Var}_4(\hat{t}_\pi) = -\frac{1}{2} \sum\sum_{sI} \frac{\Delta_{Iij}}{\pi_{Ii}} \left(\breve{t'}_i \breve{t'}_j \right) \tag{7.2.25}$$

donde $\breve{t'}_j = \frac{\hat{t}'_j}{\pi_{Ij}}$ y $\hat{t}'_j = \frac{N_j}{n'_j} \sum_{s'_j} y_k$ donde s'_j denota una muestra de n'_j elementos. La regla para determinar el n'_j y obtener el estimador $\widehat{Var}_4$ es

$$n'_j = \frac{n_i(1-\pi_{Ii})}{1-\pi_{Ii}(n_i/N_i)} \tag{7.2.26}$$

Simulación: se utilizaron los datos de la encuesta familiar de gastos FAMEX (Canada Family Expenditure, por sus siglas en inglés) del año 1996, que cuenta con un total de 691 individuos y está dividida en cinco conglomerados, se utilizó la variable gasto para estimar el total en una muestra bietápica y los datos de FAMEX 1996, aunque son los datos de una encuesta, se tomaron como los datos de un universo.

El estudio quiere verificar los resultados obtenidos anteriormente. Para el diseño de la muestra se quiso que en la primera etapa se seleccionaran tres conglomerados; para cada conglomerado seleccionado, se extrajo una muestra cuyo tamaño fuera el 40 % del mismo. El muestreo y el sub-muestreo fueron aleatorios simples MAS-MAS. El total poblacional para la variable de interés es USD 711623 y la varianza del π estimador, bajo las anteriores condiciones, es 6595944566.

Así, se calcularon los siguientes estimadores para la varianza del total estimado $\hat{t}_\pi$

- $\widehat{Var}_1(\hat{t}_\pi)$: el estimador clásico al utilizar muestreo bietápico.
- $\widehat{Var}_2(\hat{t}_\pi)$: correspondiente al primer sumando del anterior estimador.
- $\widehat{Var}_3(\hat{t}_\pi)$: el estimador suponiendo muestreo con reemplazo.

- $\widehat{Var}_4(\hat{t}_\pi)$: el estimador propuesto por (Srinath & Hidiroglou 1980) (1.5).

El proceso se repitió $B = 5000$ veces. La simulación fue programada en el paquete estadístico R. En la simulación. El desempeño de un estimador $\hat{V}$ fue evaluado usando su sesgo relativo, SR y su eficiencia relativa, ER, definidas como:

$$SR = B^{-1} \sum_{b=1}^{B} \frac{\hat{V}_b - V}{V} \tag{7.2.27}$$

$$ER = \frac{ECM(\hat{V}_\pi)}{ECM(\hat{V})} \quad , \tag{7.2.28}$$

donde

$$ECM(\hat{V}) = B^{-1} \sum_{b=1}^{B} (\hat{V}_b - V)^2 \tag{7.2.29}$$

y $\hat{V}_b$ se calculó en la b-ésima muestra simulada. Como se puede notar el estimador clásico al utilizar muestreo bietápico, $\hat{V}_\pi$, fue utilizado como línea base de comparación. Grandes valores para $ER(> 1)$ representan alta eficiencia del estimador $\hat{V}$ en comparación al estimador clásico.

$\widehat{Var}_1(\hat{t}_\pi)$	$\widehat{Var}_2(\hat{t}_\pi)$	$\widehat{Var}_3(\hat{t}_\pi)$	$\widehat{Var}_4(\hat{t}_\pi)$
0.0008138860	0.2458789480	-1.5021980054	-0.0008792021

Sesgo relativo para cada estimador

Los resultados empíricos indican que el estimador de la varianza para el estimador de Horvitz-Thompson es insesgado, así como el estimador propuesto por (Srinath & Hidiroglou 1980). Pero, los estimadores 2 y 3 tiene un sesgo relativo importante, sobre todo aquel que supone muestreo con reemplazo; también se puede observar que el estimador de la primera parte de (7.2.20), aunque es sesgado, esta magnitud es pequeña. En particular se recomienda seguir trabajando con el estimador clásico pues los avances computacionales así lo permiten. La eficiencia relativa de todos los estimadores resultó despreciable.

7.2.4 Marco II y Lucy

En el capítulo pasado se ejecutó un diseño de muestreo por conglomerados cuya principal característica es que las unidades dentro de cada conglomerado tienen un comportamiento relativamente similar. Esto llevó a que las estimaciones estuvieran muy lejos de la realidad dado que se utilizó un diseño de muestreo que inducía probabilidades de inclusión constante, siendo que el comportamiento de los totales de los conglomerados no era constante para las características de interés.

En esta oportunidad, volvemos a enfrentarnos a la dificultad de obtener una muestra de empresas del sector industrial careciendo de un marco de muestreo que nos permita la inclusión directa de las empresas en la muestra. Sin embargo, es posible utilizar como base el muestreo por áreas que se propuso en el capítulo anterior pero

la gran diferencia es que, en lugar de un censo en las áreas geográficas seleccionadas, realizaremos un sub-muestreo. Recordemos que la ciudad está dividida en cinco zonas geográficas rotuladas como **Zona A**, ubicada en el sur, **Zona B**, ubicada en el norte, **Zona C**, ubicada en el oriente, **Zona D**, ubicada en el occidente y **Zona E**, ubicada en el centro.

Suponga que no se tiene información acerca de cuántas empresas pertenecen a cada zona geográfica, por lo que no es posible realizar un diseño auto-ponderado. Para garantizar una buena precisión se ha decidido seleccionar una muestra aleatoria simple de cuatro zonas geográficas, o unidades primarias de muestreo. Lo anterior se realiza mediante el uso de la función `sample`, aunque también es admisible realizarlo con la función `S.SI`del paquete `TeachingSampling`.

```
> UI <- c("A","B","C","D","E")
> NI=length(UI)
> nI=4
> samI <- sample(NI,nI)
> muestraI <- UI[samI]
> muestraI
[1] "B" "C" "D" "E"
```

Una vez se realiza el sorteo aleatorio, las zonas geográficas seleccionadas son: **Zona B**, **Zona C**, **Zona D** y **Zona E**. El paso a seguir es el empadronamiento de cada una de las empresas del sector industrial pertenecientes a cada zona incluida en la muestra. Es decir, se debe planear un operativo de campo con el fin de levantar un marco de muestreo para cada unidad primaria. En total se deben conseguir cuatro marcos de muestreo de empresas.

```
> data(Lucy)
> attach(Lucy)

> Lucy2 <- Lucy[which(Zone==muestraI[1]),]
> Lucy3 <- Lucy[which(Zone==muestraI[2]),]
> Lucy4 <- Lucy[which(Zone==muestraI[3]),]
> Lucy5 <- Lucy[which(Zone==muestraI[4]),]

> N2 <- dim(Lucy2)[1];      n2 <- 75
> N3 <- dim(Lucy3)[1];      n3 <- 245
> N4 <- dim(Lucy4)[1];      n4 <- 50
> N5 <- dim(Lucy5)[1];      n5 <- 40

> Ni <- c(N2,N3,N4,N5)
> ni <- c(n2,n3,n4,n5)
```

Cuando la primera etapa de muestreo concluye, se tiene conocimiento de cuántas empresas del sector industrial pertenecen a cada zona geográfica incluida en la muestra. La **Zona B** con 727 empresas, la **Zona C** con 974 empresas, la **Zona D** con 223 empresas y, por último, la **Zona E** tiene un total de 165 empresas. Se ha decido que los tamaños de muestra correspondan a un porcentaje del tamaño de cada unidad primaria de muestreo. El tamaño de la muestra es de 410 empresas.

Con ayuda de cada uno de los cuatro marcos de muestreo se realiza una muestra aleatoria simple de empresas de acuerdo a los tamaños establecidos anteriormente. Cuando las muestras hayan sido seleccionadas se unifican mediante el uso de la función `rbind` que lo único que hace es mezclar las bases de datos de las empresas incluidas en la muestra.

```
> sam2 <- sample(N2,n2)
> sam3 <- sample(N3,n3)
> sam4 <- sample(N4,n4)
> sam5 <- sample(N5,n5)
>
> muestra2 <- Lucy2[sam2,]
> muestra3 <- Lucy3[sam3,]
> muestra4 <- Lucy4[sam4,]
> muestra5 <- Lucy5[sam5,]
>
> muestra <- rbind(muestra2,muestra3,muestra4,muestra5)
> attach(muestra)
> muestra
     Identificador Ubicación   Level Zona
2345        AB1206    c25k68  Grande    B
124         AB1334     c2k25 Pequeño    B
...
1204        AB2381    c13k16 Pequeño    C
713          AB187     c8k20 Pequeño    C
...
2132         AB785    c23k53 Mediano    D
2049         AB684    c22k69 Mediano    D
...
2222         AB887    c24k44 Mediano    E
2212         AB875    c24k34 Mediano    E
```

Cuando el levantamiento de la información ha concluido, se carga el archivo de datos en el ambiente de R y se construye un `data frame` que contiene los valores de las características de interés en la muestra general. En este caso particular lleva el nombre de `estima`. Es necesario que cada empresa incluída en la muestra lleve consigo el registro que indique a qué zona geográfica pertenece. Para este ejercicio, el vector `Area` contiene esta información. La estimación en este diseño de muestreo en dos etapas se hace utilizando la función `E.2SI(NI,nI,Ni,ni,y,C)` cuyos argumentos son `NI`, el número de unidades primarias de muestreo que conforman la población. `nI`, el número de unidades primarias incluidas en la muestra s_I. `Ni`, un vector de los tamaños de las unidades primarias de muestreo. `ni`, un vector conteniendo los tamaños de muestra en cada unidad primaria de muestreo. `y`, el archivo de datos que contiene la información de las características de interés y, por último, `C`, un vector que contiene la pertenencia de cada unidad secundaria de muestreo a su respectiva unidad primaria.

```
> estima <- data.frame(Income, Employees, Taxes)
> Area <- as.factor(as.integer(Zone))
> E.2SI(NI,nI,Ni,ni,estima,Area)
```

```
                   N       Income    Employees        Taxes
Total est    2136.01509 9.965934e+05 1.347360e+05 2.846822e+04
Varianza   196008.08271 1.145907e+10 3.725672e+08 4.262562e+06
cve            20.72682 1.074129e+01 1.432579e+01 7.252289e+00
```

Los resultados de la estimación se muestran en la siguiente tabla. Nótese que con un tamaño de muestra similar, la eficiencia de esta estrategia de muestreo es mucho mayor que la de una estrategia que utiliza un diseño de muestreo por conglomerados y es equivalente a la de una estrategia que utilice un diseño de muestreo aleatorio simple.

Tabla 7.1: *Muestreo aleatorio en dos etapas: estimación de los totales de las características de interés.*

Variable	Total poblacional	Total estimado	cve %	Desv. %
Ingreso	1035217	996593	1.07	-3.73
Empleados	151950	134736	1.43	-11.33
Impuestos	28654	28468	7.25	-0.65

La ganancia en eficiencia se debe a la propiedad del diseño en dos etapas en donde dado un n, es posible incluir más unidades primarias en la primera etapa de muestreo. En este caso, el número de conglomerados incluidos en la muestra s_I es el doble, lo que decrece el componente de la varianza en la primera etapa. El componente de variabilidad que domina la varianza en esta estimación es la dispersión dentro de las unidades primarias y se debe a la heterogeneidad de los conglomerados.

7.3 Muestreo en dos etapas estratificado

La teoría discutida hasta ahora en las secciones anteriores es aplicable cuando las unidades primarias de muestreo son seleccionadas de un estrato. Como se verá más adelante no hay nuevos principios de estimación o diseño involucrado en el desarrollo de esta estrategia de muestreo cuando lo que se quiere es estimar el total de la característica de interés t_y de una población dividida en H estratos.

Se supone que el muestreo en cada estrato respeta el principio de la independencia. Las estimaciones del total, así como el cálculo y estimación de la varianza son simplemente resultado de añadir o sumar para cada estrato la respectiva cantidad.

Por ejemplo, suponga que dentro de cada estrato U_h $h = 1, \ldots, H$ existen N_{Ih} unidades primarias de muestreo, de las cuales se selecciona una muestra s_{Ih} de n_{Ih} unidades mediante un diseño de muestreo aleatorio simple. Suponga, además que el sub-muestreo dentro de cada unidad primaria seleccionada es también aleatorio simple. Es decir, para cada unidad primaria de muestreo seleccionada $i \in s_{Ih}$ de tamaño N_i se selecciona una muestra s_i de elementos de tamaño n_i. Cuando las unidades secundarias de muestreo o elementos son seleccionadas, se realiza el proceso de medición y el proceso de estimación para lo cual se tiene que el estimador del total está dado por el siguiente resultado.

Resultado 7.3.1. *Bajo muestreo en dos etapas estratificado MAS-MAS, el estimador de Horvitz-Thompson es insesgado para el total poblacional y toma la*

forma

$$\hat{t}_{y,\pi}=\sum_{h=1}^{H}\hat{t}_{yh,\pi}=\sum_{h=1}^{H}\left[\frac{N_{Ih}}{n_{Ih}}\sum_{i\in S_{Ih}}\frac{N_i}{n_i}\sum_{k\in S_i}y_k\right] \tag{7.3.1}$$

con varianza dada por

$$Var_{EMM}(\hat{t}_{y,\pi})=\sum_{h=1}^{H}Var(\hat{t}_{yh,\pi}) \tag{7.3.2}$$

$$=\sum_{h=1}^{H}\left[\frac{N_{Ih}^2}{n_{Ih}}\left(1-\frac{n_{Ih}}{N_{Ih}}\right)S^2_{t_{yh}U_I}+\frac{N_{Ih}}{n_{Ih}}\sum_{i\in S_{Ih}}\frac{N_i^2}{n_i}\left(1-\frac{n_i}{N_i}\right)S^2_{y_{U_i}}\right] \tag{7.3.3}$$

cuya estimación insesgada es

$$\widehat{Var}_{EMM}(\hat{t}_{y,\pi})=\sum_{h=1}^{H}\widehat{Var}(\hat{t}_{yh,\pi}) \tag{7.3.4}$$

$$=\sum_{h=1}^{H}\left[\frac{N_{Ih}^2}{n_{Ih}}\left(1-\frac{n_{Ih}}{N_{Ih}}\right)S^2_{\hat{t}_{yh}S_I}+\frac{N_{Ih}}{n_{Ih}}\sum_{i\in S_{Ih}}\frac{N_i^2}{n_i}\left(1-\frac{n_i}{N_i}\right)S^2_{y_{S_i}}\right] \tag{7.3.5}$$

donde $S^2_{t_{yh}U_I}$ *es la varianza poblacional de los totales* t_{yi} $i\in U_I$ *de todas y cada una de las unidades primarias de muestreo dentro del estrato* h *y* $S^2_{y_{U_i}}$ *es la varianza poblacional entre los elementos dentro de cada unidad primaria de muestreo en el estrato* h*. Similarmente,* $S^2_{\hat{t}_{yh}s_I}$ *y* $S^2_{y_{s_i}}$.

Este diseño de muestreo es usado para mejorar la eficiencia de la estrategia MAS-MAS. Särndal, Swensson & Wretman (1992) plantean que es posible estratificar la población de acuerdo a una medida de tamaño, de tal forma que se agrupen las unidades de muestreo con un comportamiento similar en un mismo estrato. Es de gran interés notar que una escogencia particular dentro del sub-muestreo de las unidades primarias haría al estimador de Horvitz-Thompson muy conveniente de calcular. De hecho, si para cada unidad primaria $i\in S_{I_h}$ seleccionada en la muestra de cada estrato h, $h=1,\ldots,H$ se tiene que

$$c=\frac{n_i}{N_i}\frac{n_{Ih}}{N_{Ih}} \tag{7.3.6}$$

Entonces, el estimador toma la siguiente forma

$$\hat{t}_{y,\pi}=\frac{1}{c}\sum_{h=1}^{H}\sum_{i\in S_{Ih}}\sum_{k\in S_i}y_{hik} \tag{7.3.7}$$

Lo que significa que, en el cálculo computacional de la estimación, los valores de la característica de interés simplemente se suman sin importar la unidad primaria o el estrato al que pertenezcan. Esta clase de estimadores se conocen con el nombre de **estimadores auto-ponderados**. La cantidad c admite una interpretación muy simple y es la fracción de muestreo esperada para los elementos. De esta forma, si se desea seleccionar una muestra con un promedio de 1 % de unidades secundarias de muestreo o elementos seleccionados en cada estrato, entonces $k=\frac{1}{100}$.

7.3.1 Diseños auto-ponderados

En muchas encuestas de dos etapas es común encontrar **diseños auto-ponderados**. Esta clase de diseños asume que en la primera etapa de muestreo se selecciona una muestra S_I de unidades primarias de muestreo cuyas probabilidades de inclusión son proporcionales al tamaño de las mimas, de tal forma que si N es el tamaño de la población U de unidades secundarias de muestreo o elementos y n el tamaño de la muestra resultante, entonces

$$\pi_{Ii} = \frac{N_i}{N} n_I \qquad i \in U_I \tag{7.3.8}$$

Más adelante, en la segunda etapa de muestreo, se seleccionan muestras s_i $i \in S_I$ de unidades secundarias o elementos de tamaño constante $n_i = n_0$ para cada unidad primaria incluida en la muestra. Por lo tanto, la probabilidad de inclusión de las unidades secundarias será

$$\pi_{k|i} = \frac{n_0}{N_i} \qquad i \in S_I \tag{7.3.9}$$

De tal forma que la probabilidad de inclusión general del k-ésimo elemento es constante y está dada por

$$\pi_k = \pi_{Ii}\pi_{k|i} = n_I \frac{N_i}{N}\frac{n_0}{N_i} = n_I \frac{n_0}{N} = \frac{n}{N} = c \qquad k \in U_i \tag{7.3.10}$$

y el estimador de Horvitz-Thompson toma la siguiente forma

$$\hat{t}_{y,\pi} = \sum_{k\in S}\frac{y_k}{\pi_k} = \frac{1}{c}\sum_{i\in S_I}\sum_{k\in S_i} y_k = \frac{N}{n}\sum_{k\in S} y_k \tag{7.3.11}$$

Nótese la facilidad de cálculo del estimador. Esta clase de diseños auto-ponderados se utilizan cuando se desea controlar el trabajo de campo, por lo que el número de entrevistas en cada unidad primaria incluida en la muestra será constante.

7.4 Diseños en r etapas

Särndal, Swensson & Wretman (1992) afirman que a pesar de su complejidad, los diseños con tres o más etapas son ampliamente usados en las grandes encuestas. El muestreo en dos etapas puede ser generalizado mediante el siguiente resultado en donde se supone que existen r etapas de muestreo. De esta manera, la población se divide en N_I unidades primarias de muestreo, de las cuales se selecciona una muestra s_I de n_I unidades mediante un diseño de muestreo $p_I(S_I)$. Se asume que es posible construir un estimador[3] $\hat{t}_{yi}$ para cada total t_{yi} $i \in S_I$ de las unidades primarias seleccionadas y que este estimador es insesgado para las restantes $r-1$ etapas del diseño muestral. Por tanto

[3]Este estimador no necesariamente debe ser el estimador de Horvitz-Thompson pero sí debe ser insesgado.

$$E(\hat{t}_{yi} \mid S_I) = t_{yi} \tag{7.4.1}$$

Nótese que las últimas unidades de muestreo no deben ser necesariamente elementos, pueden ser también conglomerados. Los principios de independencia e invarianza se siguen manteniendo en todas las etapas del diseño muestral. De tal manera que el fundamento de este diseño de muestreo es la acumulación de las estimaciones desde la última etapa hasta la primera. Esto se sintetiza en los siguientes resultado de la próxima sección.

7.4.1 El estimador de Horvitz-Thompson

Resultado 7.4.1. *Bajo muestreo en r etapas el estimador de Horvitz-Thompson es insesgado para el total poblacional y toma la forma*

$$\hat{t}_{y,\pi} = \sum_{i \in S_I} \frac{\hat{t}_{yi}}{\pi_{Ii}} \tag{7.4.2}$$

con varianza dada por

$$Var_{BI}(\hat{t}_{y,\pi}) = \underbrace{\sum\sum_{U_I} \Delta_{Iij} \frac{t_i}{\pi_{Ii}} \frac{t_j}{\pi_{Ij}}}_{Var(UPM)} + \underbrace{\sum_{i \in U_I} \frac{V_i}{\pi_{Ii}}}_{Var(Resto)} \tag{7.4.3}$$

cuya estimación insesgada es

$$\widehat{Var}_{BI}(\hat{t}_{y,\pi}) = \underbrace{\sum\sum_{S_I} \frac{\Delta_{Iij}}{\pi_{Iij}} \frac{\hat{t}_{yi}}{\pi_{Ii}} \frac{\hat{t}_{yj}}{\pi_{Ij}}}_{\widehat{Var}(UPM)} + \underbrace{\sum_{i \in S_I} \frac{\hat{V}_i}{\pi_{Ii}}}_{\widehat{Var}(Resto)} \tag{7.4.4}$$

donde $V_i = Var(\hat{t}_{yi} \mid S_I)$ y $\hat{V}_i$ es un estimador insesgado de V_i tal que $E(\hat{V}_i \mid S_I) = V_i$ para todo $i \in U_I$.

Prueba. Esta demostración se realiza de manera recursiva escribiendo el estimador y la varianza como una función de los estimadores insesgados de las etapas subsecuentes en los niveles inferiores. Se debe tener en cuenta que el resultado 7.2.2. se extiende naturalmente. Por ejemplo para el diseño de tres etapas, se tiene que

$$Var(U) = V_1[E_2(E_3(U))] + E_1[V_2(E_3(U))] + E_1[E_2(V_3(U))] \tag{7.4.5}$$

■

7.4.2 El estimador de Hansen-Hurwitz

Un esquema utilizado en la práctica por la sencillez en el proceso de estimación consiste en seleccionar una muestra de m_I unidades primarias de muestreo mediante un diseño de muestreo con reemplazo que induce probabilidades de selección p_{Ii}

con $i \in U_I$ tales que $\sum_{i=1}^{N_i} p_{Ii} = 1$. Dentro de cada unidad primaria de muestreo seleccionada en el sorteo aleatorio con reemplazo se toma una sub-muestra (con o sin reemplazo). Aunque existe una pérdida de eficiencia cuando el muestreo es con reemplazo, ésta se compensa con una ganancia logística en el proceso de estimación de las varianzas requeridas para cada característica de interés. El proceso general de muestreo con reemplazo según Särndal, Swensson & Wretman (1992) es el siguiente:

- En la primera etapa se selecciona una muestra aleatoria de acuerdo a un diseño de muestreo con reemplazo tal que p_{Ii} con $i \in U_I$ es la probabilidad de selección de la i-ésima unidad primaria de muestreo.
- En las siguientes etapas[4], se mantienen las propiedades de independencia e invarianza sin importar si el diseño dentro de las unidades primarias seleccionadas sea con o sin reemplazo.
- Si una unidad de muestreo es seleccionada en más de una ocasión, se debe realizar tantos sub-muestreos como veces haya sido seleccionada en la primera etapa.

Resultado 7.4.2. *Bajo un diseño de muestreo en varias etapas, el estimador de Hansen-Hurwitz para el total t_y, su varianza y su varianza estimada están dados por*

$$\hat{t}_{y,p} = \frac{1}{m_I}\sum_{v=1}^{m_I}\frac{\hat{t}_{yi_v}}{p_{Ii_v}} \tag{7.4.6}$$

$$Var(\hat{t}_{y,p}) = \frac{1}{m_I}\sum_{i=1}^{N_I} p_{Ii}\left(\frac{t_{yi}}{p_{Ii}} - t_y\right)^2 + \frac{1}{m_I}\sum_{i=1}^{N_I}\frac{V_i}{p_{Ii}} \tag{7.4.7}$$

$$\widehat{Var}(\hat{t}_{y,p}) = \frac{1}{m_I(m_I-1)}\sum_{v=1}^{m_I}\left(\frac{\hat{t}_{yi_v}}{p_{Ii_v}} - \hat{t}_{y,p}\right)^2 \tag{7.4.8}$$

respectivamente. Donde $\hat{t}_{yi}$ es un estimador insesgado del total de la característica de interés y en la unidad primaria U_i $i \in S_I$, $V_i = Var(\hat{t}_{yi} \mid S_I)$ la varianza de $\hat{t}_{yi}$ en la segunda etapa. Nótese que $\hat{t}_{y,p}$ es insesgado para t_y y que $\widehat{Var}(\hat{t}_{y,p})$ es insesgado para $Var(\hat{t}_{y,p})$.

Prueba. La demostración empieza definiendo las variables aleatorias

$$Z_v = t_{yi}/p_{Ii} \qquad i \in U_I \quad v = 1, \ldots, m_I \tag{7.4.9}$$

y

$$\hat{Z}_v = \hat{t}_{yi}/p_{Ii} \qquad i \in U_I \quad v = 1, \ldots, m_I \tag{7.4.10}$$

Tanto Z_v como $\hat{Z}_v$ son sucesiones de variables aleatorias independientes e idénticamente distribuidas. Sin embargo, respetando los principios de independencia e invarianza, se tiene que la esperanza está dada por

$$E(\hat{Z}_v) = E(E(\hat{Z}_v \mid S_I)) = E(Z_v) = t_y$$

[4]Este proceso es válido para diseños de muestreo con más de dos etapas.

y la varianza es

$$\begin{aligned}
Var(\hat{Z}_v) &= Var(E(\hat{Z}_v \mid S_I)) + E(Var(\hat{Z}_v \mid S_I)) \\
&= Var(Z_v) + E(Var(\hat{t}_{yi}/p_{Ii} \mid S_I)) \\
&= Var(Z_v) + E(V_i/p_{Ii}^2) \\
&= \sum_{i=1}^{N_I} p_{Ii}\left(\frac{t_{yi}}{p_{Ii}} - t_y\right)^2 + \sum_{i=1}^{N_I} \frac{V_i}{p_{Ii}}
\end{aligned}$$

Ahora, dado que $\hat{t}_{y,p} = \overline{\hat{Z}}$ y utilizando el resultado 2.2.11, se tiene que el estimador insesgado de la varianza corresponde a la expresión dada en (7.4.8). ■

Dada la simplificación en el cálculo de la varianza, Bautista (1998) propone utilizarla incluso cuando el diseño de muestreo sea sin reemplazo. Sin embargo, advierte que este estimador generalmente sobre-estima la varianza, lo que conduce a intervalos de confianza más conservadores y coeficientes de variación un poco más altos.

7.5 Ejercicios

7.1 Argumente si las siguientes afirmaciones son falsas o verderas. Sustente su respuesta detallamente.

(a) En la estimación de totales poblaciones, se nota que, casi siempre, $Var_{MAS^2}(t_{y,\pi})$ es mayor a $Var_{MAS}(t_{y,\pi})$.

(b) En la estimación de la varianza para totales en diseños bietápicos, $\hat{Var}(UPM)$ es insesgada para $\hat{Var}(UPM)$.

(c) En la estimación de la varianza para totales en diseños bietápicos, $\hat{Var}(USM)$ es insesgada para $\hat{Var}(USM)$.

(d) Al planear un diseño de muestreo en varias etapas, se debe tener en cuenta que entre más etapas tenga el diseño, la varianza del estimador será probablemente más baja.

(e) En diseños bietápicos, la varianza total del estimador es dominada por la varianza de la última etapa. Es decir, la varianza en la última etapa es mucho mayor que la varianza de la primera etapa.

(f) En un estudio de consumo de licores se proponen dos diseños de muestreo en dos etapas: uno con la selección de 300 manzanas y diez personas por manzana; el otro con la selección de 100 manzanas y 30 personas por manzana. En este caso, el primer diseño de muestreo arroja una varianza menor al del segundo diseño.

7.2 Para un diseño de muestreo en dos etapas, en donde la primera etapa se lleva a cabo un diseño PPT con reemplazo y en la segunda etapa se realiza un diseño MAS en cada UPM seleccionada, proponga un estimador insesgado para el total poblacional (Ayuda: utilice el estimador de Horvitz-Thompson

en la segunda etapa y el estimador de Hansen-Hurwitz en la primera etapa). Demuestre que este estimador es insesgado para el total poblacional t_y (Ayuda: utilice las propiedades de la esperanza condicional) y defina la varianza para este estimador (Ayuda: utilice las propiedades de la varianza condicional).

7.3 Escriba las fórmulas del estimador del total y del estimador de la varianza del total para los siguientes diseños de muestreo. Defina estrictamente cada término y notación que utilice en las fórmulas.

(a) Diseño en tres etapas: MAS en cada una de las etapas.

(b) Diseño estratificado con tres estratos: uno de inclusión forzosa, otro con diseño PPT y otro con diseño MAS.

7.4 (Tillé, 2006. Ej 5.5) Suponga que un estadístico desea estimar el ingreso total de las personas en un país. Para esto, él lleva a cabo un diseño de muestreo en dos etapas, en donde la primera etapa se seleccionan municipios con un diseño PPT con probabilidad de selección proporcional al número de habitantes del municipio y en la segunda etapa se realiza un diseño MAS en cada municipio. En la primera etapa, se seleccionaron $m_I = 4$ municipios entre los $N_I = 30$ municipios en el país y en la segunda etapa, se incluyeron n_i personas de los N_i habitantes del municipio i-ésimo ($i = 1, 2, 3, 4$). Suponga que por fuentes oficiales, se conoce que el número total de personas en el país es de $N = 10000$. Los datos obtenidos se muestran en la tabla 7.2.

Tabla 7.2: Ingreso de cada persona para el ejercicio 7.3

Municipio	Ni	ni	yk
1	20	4	105 118 102 110
2	23	5	108 117 134 108 119
1	18	4	201 201 210 206
2	28	6	157 141 129 170 104 110

(a) Estime el ingreso total en el pais. Reporte el coeficiente de variación estimado.

(b) Estime el ingreso medio en el pais y reporte el coeficiente de variación estimado.

7.5 Suponga que por alguna circunstancia, un extraterrestre desea estimar el número promedio de patas que tiene un perro en una ciudad. La ciudad está dividida en dos áreas geográficas, la zona norte y la zona sur. Para llevar a cabo la estimación, él planea un diseño de muestreo en dos etapas así: De las $N_I = 2$ zonas geográficas de la ciudad, va a seleccionar una muestra aleatoria simple de $n_I = 1$ unidades primarias de muestreo. Se sabe que en el norte hay $N_1 = 30$ perros y en el sur hay $N_2 = 10$ perros. Sea cual sea la unidad primaria seleccionada, se seleccionará una sub-muestra aleatoria simple de $n_i = 2$ perros ($i = 1, 2$) y se realizará la medición del total de patas en cada perro incluido en la muestra.

(a) Si se seleccionó la zona norte, reporte la estimación del total de patas en la ciudad $t_{y,\pi}$ y la estimación del promedio de patas en la ciudad $\bar{y}_S = t_{y,\pi}/N$.

(b) Si se seleccionó la zona sur, reporte la estimación del total de patas en la ciudad $t_{y,\pi}$ y la estimación del promedio de patas en la ciudad $\bar{y}_S = t_{y,\pi}/N$.

(c) Para este diseño diseño de muestreo, reporte la varianza teórica del estimador $\bar{y}_S$.

(d) ¿Es una buena estrategia escoger al estimador $\bar{y}_S$ para inferir acerca del promedio de patas de los perros en la ciudad?

Parte II

Inferencia asistida por modelos y el uso de la información auxiliar

Capítulo 8

Estimación de parámetros diferentes al total

> Naturalmente, el investigador está interesado en encontrar las propiedades estadísticas de un estimador. Si éste tiene una forma lineal, no se necesitan nuevas herramientas. Sin embargo, los parámetros que se encuentran en la práctica corresponden a funciones no lineales de totales.
>
> Carl-Erik Särndal (1992)

En los capítulos anteriores, nuestra atención estuvo centrada en la búsqueda del mejor diseño de muestreo con los estimadores de Horvitz-Thompson, para muestreo sin reemplazo y estimadores de Hansen-Hurwitz, para muestreo con reemplazo. En nuestra travesía hemos pasado por los diseños de probabilidad fija e igual. Para mejorar la eficiencia de la estrategia hemos revisado los diseños de probabilidades proporcionales y diseños estratificados, con la ayuda de información auxiliar de tipo continuo o discreto. Para mejorar la eficacia del plan operativo y la dispersión de la muestra en la población se han propuesto diseños de muestreo complejos de conglomerados y en varias etapas.

El lector debió notar que en la primera parte de este texto se ha seguido con fidelidad la regla de oro del diseño de encuestas y es utilizar estrategias de muestreo que induzcan probabilidades de inclusión o selección, según sea el caso, proporcionales al valor de la característica de interés. De este modo, si la encuesta está enfocada en una característica de interés cuya dispersión es muy baja, como el número de hijos en niveles socioeconómicos altos, que generalmente no es mayor a tres, es posible utilizar un muestreo aleatorio con probabilidades simples. De otra manera y con la ayuda de información auxiliar, es posible seguir la regla de oro mediante la construcción de probabilidades proporcionales en la etapa de diseño. Sin embargo, esta ventaja del marco de muestreo no sólo se puede utilizar en la etapa de diseño sino también en la etapa de estimación.

8.1 Fundamentos teóricos

Siguiendo la filosofía del título que lleva este texto, nos encaminaremos en la búsqueda de la mejor estrategia de muestreo mejorando el estimador. En esta etapa del camino, se supone que el lector conoce el comportamiento estructural de la población y está en capacidad de proponer el mejor diseño de muestreo, de acuerdo a la generosidad del marco de muestreo.

Por supuesto, en algunos estudios multi-propósito, en encuestas complejas y en casos particulares, es necesario obtener estimaciones para parámetros diferentes a los totales. Por ejemplo, razones de dos características de interés, medianas y percentiles poblacionales, parámetros de regresión, coeficientes de correlación, varianzas, covarianzas, índices, etc. Como lo afirma Bautista (1998), la metodología que se propone para estimar estos parámetros poblacionales es reescribirlos como función de totales poblacionales. Así, si el parámetro a estimar es B, lo debemos llevar a la siguiente forma

$$B = f(t_1, t_2, \ldots, t_Q) \tag{8.1.1}$$

Donde cada t_q $q = 1, \ldots, Q$ representa un total de las características de interés o un total de una función de las características de interés. El principio de estimación de este parámetro está en obtener estimadores insesgados $\hat{t}_q$ $q = 1, \ldots, Q$ tal que T es estimado por

$$\hat{B} = f(\hat{t}_1, \hat{t}_2, \ldots, \hat{t}_Q) \tag{8.1.2}$$

Nótese que la función f puede ser lineal o no. Un resultado muy conocido de la inferencia estadística clásica nos indica que si la función f es una función lineal entonces B toma la forma

$$B = a_0 + \sum_{q=1}^{Q} a_q t_q \tag{8.1.3}$$

Por tanto, un estimador insesgado de B está dado por la siguiente expresión

$$\hat{B} = a_0 + \sum_{q=1}^{Q} a_q \hat{t}_q \tag{8.1.4}$$

Si en la estimación de B hemos utilizado estimadores de tipo Horvitz-Thompson, entonces es posible escribir (8.1.3) como

$$\hat{B}_\pi = a_0 + \sum_{k \in S} \frac{E_k}{\pi_k} \tag{8.1.5}$$

donde $E_k = \sum_{q=1}^{Q} a_q y_{qk}$ y el valor del k-ésimo elemento en la q-ésima característica de interés está dado por y_{jk}. Siguiendo los principios del estimador de Horvitz-Thompson, la varianza de $\hat{B}_\pi$ se puede expresar como

$$Var(\hat{B}_\pi) = \sum\sum_U \Delta_{kl} \frac{E_k}{\pi_k} \frac{E_l}{\pi_l}. \tag{8.1.6}$$

Un estimador insesgado para la expresión (8.0.5) está dada por

$$\widehat{Var}_1(\hat{B}_\pi) = \sum\sum_S \frac{\Delta_{kl}}{\pi_{kl}} \frac{E_k}{\pi_k} \frac{E_l}{\pi_l} \tag{8.1.7}$$

Nótese que cuando la función f es lineal no se involucran nuevos principios de estimación. Por el contrario, cuando f no es lineal, el estimador propuesto es la misma expresión (8.1.2); sin embargo, en algunos casos, no es posible ni calcular, ni estimar la varianza debido a la complejidad matemática teórica del desarrollo y es necesario recurrir a métodos que permitan llegar a una expresión que aproxime la varianza. Es posible aproximar la varianza utilizando las técnicas de linealización para estimar la precisión de estos estimadores. Éstas han sido introducida por Woodruff (1971). Algunas aplicaciones en la teoría de muestreo han sido desarrolladas, entre otros, por Binder (1983) y Deville (1999). El método más común, aunque no el único, es el de linealización por polinomios de Taylor.

8.1.1 Aproximación de una función por polinomios

En Apostol (1963, p. 417) se presentan las condiciones para que una función f se pueda aproximar mediante un polinomio. Entre ellas tenemos que la función f sea derivable y que sus derivadas deben estar definidas en el punto $x = a$.

Resultado 8.1.1 (Teorema de Taylor). *Si una función se puede aproximar mediante un polinomio, entonces éste estará definido por*

$$f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \ldots + \frac{f^{(n)}}{n!}(x-a)^n + \ldots \tag{8.1.8}$$

Prueba. Sea

$$f(x) = c_0 + c_1(x-a) + c_2(x-a)^2 + \ldots \tag{8.1.9}$$

Derivando sucesivamente, tenemos

$$\begin{aligned}
f^{(1)}(x) &= c_1 + 2c_2(x-a) + 3(x-a)^2 + \ldots \\
f^{(2)}(x) &= 2c_2 + 6c_3(x-a) + 12c_4(x-a)^2 + \ldots \\
f^{(3)}(x) &= 6c_3 + 24c_4(x-a) + 60c_5(x-a)^2 + \ldots \\
&\vdots \\
f^{(n)}(x) &= n!c_n + (n+1)!C_{n+1}(x-a) + (n+2)!C_{n+2}(x-a)^2 + \ldots
\end{aligned}$$

Haciendo $x = a$ tenemos

$$\begin{aligned}
f(a) &= c_0 & f^{(1)}(a) &= c_1 \\
f^{(2)}(a) &= 2c_2 & f^{(3)}(a) &= 6c_3
\end{aligned}$$

y en general $f^{(n)}(x) = n!c_n$. Sustituyendo en (8.1.9), se llega a la aproximación mediante polinomios de Taylor como en (8.1.8). ■

Para funciones vectoriales, existe el siguiente teorema de Taylor

Resultado 8.1.2. *Para una función vectorial f, se tiene que la aproximación de Taylor de primer orden de la función f en un punto (vectorial)* **a** *está dada por*

$$f(\mathbf{x}) \cong f(\mathbf{a}) + (\bigtriangledown f|_{\mathbf{x}=\mathbf{a}})'(\mathbf{x}-\mathbf{a}), \tag{8.1.10}$$

con $\mathbf{x} = (x_1, \cdots, x_Q)'$ y $\bigtriangledown f$ denota el gradiente de la función f; esto es, el q-ésimo componente de $\bigtriangledown f$ está dado por

$$\frac{\partial f(x_1, \cdots, x_Q)}{\partial x_q}.$$

Ejemplo 8.1.1. Es posible representar a la función $\sin(x)$ en series de potencias de x (es decir en el punto $a = 0$). Para este caso particular se tiene que:

$$\begin{aligned}
f(x) &= \sin(x) & f(0) &= 0\\
f^{(1)}(x) &= \cos(x) & f^{(1)}(0) &= 1\\
f^{(2)}(x) &= -\sin(x) & f^{(2)}(0) &= 0\\
f^{(3)}(x) &= -\cos(x) & f^{(3)}(0) &= -1\\
f^{(4)}(x) &= \sin(x) & f^{(4)}(0) &= 0\\
&\vdots & &\vdots
\end{aligned}$$

Por tanto, el desarrollo de la función en series es de la siguiente manera:

$$\begin{aligned}
\sin(x) &= 0 + x + \frac{0}{2!}x^2 + \frac{-1}{3!}x^3 + \frac{0}{4!}x^4 + \frac{1}{5!}x^5 + \ldots\\
&= x + \frac{-1}{3!}x^3 + \frac{1}{5!}x^5 + \ldots\\
&= \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{(2n-1)!}x^{(2n-1)}
\end{aligned}$$

Sin embargo, no solamente debemos revisar si la función y sus derivadas están definidas en un punto $x = a$, también debemos revisar la convergencia de la serie de potencias. Para esto utilizaremos la prueba de convergencia de la razón definido en Apostol (1963, p. 363). Esta prueba argumenta que si el resultado de R, definido por

$$R = \lim_{n\to\infty} \left|\frac{S_{n+1}}{S_n}\right|, \tag{8.1.11}$$

es menor que uno, entonces la serie converge absolutamente. Para este ejemplo particular, tenemos que

$$
\begin{aligned}
R &= \lim_{n\to\infty}\left|\frac{(-1)^{(n-1)+1}x^{2(n+1)-1}}{(2(n+1)-1)!}\Bigg/\frac{(-1)^{n-1}x^{2n-1}}{(2n-1)!}\right| \\
&= \lim_{n\to\infty}\left|\frac{x^{2n+1}}{(2n+1)!}\frac{(2n-1)!}{x^{2n-1}}\right| \\
&= x^2\lim_{n\to\infty}\left|\frac{1}{2n(2n+1)}\right| = 0
\end{aligned}
$$

Por lo tanto, la serie converge absolutamente y tendríamos una buena aproximación a $f(x) = sin(x)$ al cortar la serie y dejar un residuo que sería despreciable.

Aplicación en muestreo

Mediante esta técnica es posible aproximar la varianza de los estimadores que no son funciones lineales de totales. Aunque en el ámbito de la inferencia en poblaciones finitas, no existe una teoría asintótica unificada, sí existen resultados particulares para los diseños de muestreo más simples Madow (1948) y para algunos diseños de muestreo con probabilidades proporcionales Rosén (1972). Lohr (2000) plantea los siguientes pasos para construir un estimador linealizado de la varianza de una función no lineal de totales:

1. Expresar el estimador del parámetro de interés $\hat{B}$ como una función de estimadores de totales insesgados. Así, $\hat{B} = f(\hat{t}_1, \hat{t}_2, \ldots, \hat{t}_Q)$.

2. Determinar todas las derivadas parciales de f con respecto a cada total estimado $\hat{t}_{q,\pi}$ y evaluar el resultado en las cantidades poblacionales t_q. Así

$$a_q = \left.\frac{\partial f(\hat{t}_1, \ldots, \hat{t}_Q)}{\partial \hat{t}_q}\right|_{\hat{t}_1=t_1,\ldots,\hat{t}_Q=t_Q} \tag{8.1.12}$$

3. Aplicar el teorema de Taylor para funciones vectoriales para linealizar la estimación $\hat{B}$ con $\mathbf{a} = (t_1, t_2, \cdots, t_Q)'$. En el paso anterior, se vio que $\bigtriangledown\hat{B}' = (a_1, \cdots, a_Q)$. Por consiguiente se tiene que

$$\hat{B} = f(\hat{t}_1, \ldots, \hat{t}_Q) \cong B + \sum_{q=1}^{Q} a_q(\hat{t}_q - t_q) \tag{8.1.13}$$

4. Definir una nueva variable E_k con $k \in S$ al nivel de cada elemento observado en la muestra aleatoria.

$$E_k = \sum_{q=1}^{Q} a_q y_{qk} \tag{8.1.14}$$

5. De (8.1.12) y (8.1.13) se tiene que, si los estimadores $\hat{t}_q$ son estimadores de Horvitz-Thompson, una expresión que aproxima la varianza de $\hat{B}$ está dada

por

$$AVar(\hat{B}) = Var\left(\sum_{q=1}^{Q} a_q \hat{t}_{q,\pi}\right)$$
$$= Var\left(\sum_{S} \frac{E_k}{\pi_k}\right) = \sum\sum_{U} \Delta_{kl} \frac{E_k}{\pi_k} \frac{E_l}{\pi_l}.$$

Para encontrar una estimación de la varianza de $\hat{B}$, no es posible utilizar directamente los valores E_k, porque éstos dependen de los totales poblacionales, pues las derivadas a_q se evalúan en los totales poblacionales que son desconocidos. Por consiguiente, los valores E_k se aproximan reemplazando los totales desconocidos por los estimadores de los mismos. Siendo e_k la aproximación de la variable linealizada dada por

$$e_k = \sum_{q=1}^{Q} \hat{a}_q y_{qk} \tag{8.1.15}$$

donde $\hat{a}_q$ corresponde a un estimador de a_q. Por otro lado, Deville (1999) ha probado que la aproximación de la varianza lograda mediante e_k es válida para grandes tamaños de muestra. Si los estimadores $\hat{t}_q$ son estimadores de Horvitz-Thompson, se puede usar de manera general el estimador de la varianza de Horvitz-Thompson, así

$$\widehat{Var}(\hat{t}_{y,\pi}) = \sum\sum_{S} \frac{\Delta_{kl}}{\pi_{kl}} \frac{e_k}{\pi_k} \frac{e_l}{\pi_l} \tag{8.1.16}$$

Como siempre, si el diseño de muestreo es de tamaño fijo, se pueden utilizar las respectivas expresiones dadas en el capítulo 2 de este texto. Särndal, Swensson & Wretman (1992) advierten que este método tiende a sub-estimar la varianza real cuando el tamaño de muestra es pequeño. Por otra parte, una desventaja de este método es la particularidad de cada aproximación sujeta a la forma funcional del parámetro de interés. De esta manera, es necesario determinar expresiones analíticas particulares. Esto genera desgaste cuando se trabaja con encuestas complejas. El siguiente resultado resume el proceso de inferencia general para la estimación de una función linealizada de totales.

Resultado 8.1.3. *Siendo $B = f(t_1, t_2, \ldots, t_Q)$ es una función de totales poblacionales, entonces un estimador aproximadamente insesgado de B, su varianza aproximada y una estimación insesgada para esta última están dadas por las siguientes expresiones*

$$\hat{B}_\pi = f(\hat{t}_{1,\pi}, \hat{t}_{2,\pi}, \ldots, \hat{t}_{Q,\pi}) \tag{8.1.17}$$

$$AVar(\hat{B}_\pi) = \sum\sum_{U} \Delta_{kl} \frac{E_k}{\pi_k} \frac{E_l}{\pi_l} \tag{8.1.18}$$

$$\widehat{Var}(\hat{B}_\pi) = \sum\sum_{S} \frac{\Delta_{kl}}{\pi_{kl}} \frac{e_k}{\pi_k} \frac{e_l}{\pi_l} \tag{8.1.19}$$

respectivamente, con $\hat{t}_{q,\pi}$ el estimador de Horvitz-Thompson de $t_{q,\pi}$ y tanto E_k como e_k se encuentran dados por las fórmulas (8.1.14) y (8.1.15), en estricto orden.

Prueba. En primer lugar,

$$
\begin{aligned}
E(\hat{B}_\pi) &\cong E(B + \sum_{q=1}^{Q} a_q(\hat{t}_q - t_q)) \\
&= B + \sum_{q=1}^{Q} a_q E(\hat{t}_q - t_q) \\
&= B
\end{aligned}
$$

puesto que $\hat{t}_q$ es insesgado para t_q, para $q = 1, \cdots, Q$. Por otro lado,

$$
\begin{aligned}
Var(\hat{B}_\pi) &= Var(\sum_{q=1}^{Q} a_q \hat{t}_q) \\
&= Var(\sum_{q=1}^{Q} a_q \sum_{k \in S} \frac{y_{qk}}{\pi_k}) \\
&= Var(\sum_{k \in S} \frac{E_k}{\pi_k}) \\
&= \sum\sum_{U} \Delta_{kl} \frac{E_k}{\pi_k} \frac{E_l}{\pi_l}
\end{aligned}
$$

■

8.2 Estimación de una razón poblacional

Un caso especial de una función no-lineal de totales es la razón poblacional B. Ésta se define como el cociente de dos totales poblacionales de características de interés z e y. Así

$$
B = \frac{t_y}{t_z} = \frac{\bar{y}_U}{\bar{z}_U} \tag{8.2.1}
$$

Lohr (2000) plantea que técnicamente siempre se estimará una razón cuando se estime un promedio de un dominio. Nótese que la característica de la razón es que tanto el denominador como el numerador son desconocidos, y aunque se conocieran, se prefieren estimar. Bautista (1998) da ejemplos muy concretos en lo que se utilizó la estimación de razones. Entre ellos están los siguientes:

- **Estudios electorales:** para estimar la intención de voto por un candidato se pregunta por qué candidato votaría el encuestado[1]. Dado que no todas las personas entrevistadas pueden votar, incluso algunos de ellos decidirán no votar por omisión. El numerador de esta razón está dado por el total de personas que votarían por el candidato, mientras que el denominador de la razón sería el total de personas que participarían activamente en las elecciones. Nótese que la tasa de abstención también está dada por una razón.

[1]Bajo el supuesto de que las elecciones se realizarían el mismo día de la entrevista.

El numerador correspondería al total de personas que, sin tener restricción alguna, han decidido no participar en las elecciones. El denominador estaría dado por el total de personas que están aptas para votar.

- **Investigación de medios:** es importante para los canales de televisión tener un estimativo del total de personas observan algún programa de televisión en determinado momento. Con esta información, los canales cobran más o menos dinero a las empresas que deseen pautar un comercial a determinada hora. Si el programa televisivo tiene una audiencia alta, el canal cobrará más por la pauta de un comercial. Para estandarizar esta información, se ha creado un índice llamado «rating» que se define como la razón entre el total de personas que están observando un programa de televisión en un minuto determinado sobre el total de personas que están observando televisión.

- **Investigación social:** uno de los indicadores económicos que más llama la atención en el desarrollo de una región o país es la tasa de desempleo. Hay que tener en cuenta que no todos los habitantes de una región están aptos para trabajar, pues existe un rango de edad para ello. Este indicador económico está definido como el total poblacional de personas que se encuentran en edad laboral pero que carecen de un empleo sobre la cantidad de personas que pertenecen a la población económicamente activa.

Para la estimación de razones se propone el siguiente resultado que da cuenta de las expresiones teóricas que deben utilizarse para tal fin.

Resultado 8.2.1. *Un estimador para la razón poblacional B de dos características de interés, su varianza y su varianza estimada están dados por*

$$\hat{B} = \frac{\hat{t}_{y,\pi}}{\hat{t}_{z,\pi}} \tag{8.2.2}$$

$$AVar(\hat{T}_\pi) = \sum\sum_U \Delta_{kl} \frac{E_k}{\pi_k} \frac{E_l}{\pi_l}. \tag{8.2.3}$$

$$\widehat{Var}(\hat{t}_{y,\pi}) = \sum\sum_S \frac{\Delta_{kl}}{\pi_{kl}} \frac{e_k}{\pi_k} \frac{e_l}{\pi_l} \tag{8.2.4}$$

donde $E_k = \frac{1}{t_x}(y_k - Bz_k)$ y $e_k = \frac{1}{\hat{t}_{z,\pi}}(y_k - \hat{B}z_k)$ Nótese que $\hat{B}$ es aproximadamente insesgado para B al igual que $\widehat{Var}(\hat{t}_{y,\pi})$ lo es para $AVar(\hat{t}_{y,\pi})$

Prueba. Siguiendo los pasos de linealización de la sección anterior tenemos que el estimador propuesto es una función de dos totales estimados de las características de interés

$$\hat{B} = \frac{\hat{t}_{y,\pi}}{\hat{t}_{z,\pi}} = f(\hat{t}_{y,\pi}, \hat{t}_{z,\pi})$$

Calculando las derivadas parciales

$$\begin{aligned}a_1 &= \frac{\partial f(\hat{t}_{y,\pi}, \hat{t}_{z,\pi})}{\partial \hat{t}_{y,\pi}}\bigg|_{\hat{t}_{y,\pi}=t_y, \hat{t}_{z,\pi}=t_z}\\ &= \frac{1}{t_z}\\ a_2 &= \frac{\partial f(\hat{t}_{y,\pi}, \hat{t}_{z,\pi})}{\partial \hat{t}_{z,\pi}}\bigg|_{\hat{t}_{y,\pi}=t_y, \hat{t}_{z,\pi}=t_z}\\ &= -\frac{t_y}{t_z^2}\end{aligned}$$

Utilizando la aproximación de la razón mediante la expresión (8.1.12) se tiene que

$$\hat{B} = B + \frac{1}{t_z}(\hat{t}_{y,\pi} - t_y) - \frac{t_y}{t_z^2}(\hat{t}_{z,\pi} - t_z)$$

por tanto al evaluar la esperanza se tiene inmediatamente la propiedad del insesgamiento aproximado. Por otro lado, definiendo la nueva variable linealizada dada en (8.1.14), tenemos que

$$E_k = \frac{y_k}{t_z} - \frac{t_y}{t_z^2} z_k = \frac{1}{t_z}(y_k - Bz_k) \tag{8.2.5}$$

cuya aproximación es

$$e_k = \frac{1}{\hat{t}_{z,\pi}}(y_k - \hat{B}z_k) \tag{8.2.6}$$

Por tanto la varianza se escribe como

$$AVar(\hat{B}) = Var\left(\sum_S \frac{E_k}{\pi_k}\right) \tag{8.2.7}$$

Utilizando los principios del estimador de Horvitz-Thompson se llega a los resultados de la aproximación de la varianza y de la varianza estimada. ■

No es difícil probar que cualquiera que sea el diseño de muestreo utilizado siempre se cumplen las siguientes condiciones

$$\sum_U E_k = 0 \tag{8.2.8}$$

$$\sum_S \frac{e_k}{\pi_k} = 0 \tag{8.2.9}$$

8.2.1 Propiedades

Aunque la característica del insesgamiento es deseada en los estimadores, no se debe exagerar descartando algunos estimadores que tengan un poco de sesgo. En algunos casos la forma funcional del parámetro de interés es tan compleja que resulta muy complicado obtener un estimador exactamente insesgado. Por otro lado, puede existir un estimador con poco sesgo y con menor error cuadrático medio que

un estimador insesgado. De hecho, Särndal, Swensson & Wretman (1992) afirman que son muchos los estimadores aproximadamente insesgados que se utilizan en la práctica. También afirma que se debe mantener siempre presente la regla de Hájek que proclama que:

> Los estimadores con un sesgo considerable son pobres sin importar qué otras propiedades puedan tener.

Como esta clase de estimadores son aproximadamente insesgados, es necesario evaluar otro tipo de bondades como la consistencia dada en la siguiente definición.

Definición 8.2.1. *Un estimador $\hat{T}$ es consistente en el sentido Cochran para un parámetro de interés T si $s = U$ implica que el estimador reproduce el parámetro de interés. Es decir $\hat{T} = T$.*

Nótese que bajo la clase de diseños MAS, el estimador de Horvitz-Thompson es consistente pues si $s = U$, entonces $\pi_k = 1$, por lo tanto

$$\hat{t}_{y,\pi} = \sum_{k\in s} \frac{y_k}{\pi_k} = \sum_{k\in U} y_k = t_y \tag{8.2.10}$$

Sin embargo, bajo el diseño de Bernoulli, el estimador de Horvitz-Thompson no conserva la propiedad de consistencia. Suponga que las probabilidades de inclusión de primer orden están dadas por $\pi = 0.1$. El evento $s = U$ ocurre con probabilidad 0.1^N, para el cual el estimador de Horvitz-Thompson tomaría la siguiente forma

$$\hat{t}_{y,\pi} = \sum_{k\in s} \frac{y_k}{0.1} = 10 \times t_y \tag{8.2.11}$$

Nótese que bajo este escenario, él estimador de razón $\hat{B}$ es consistente.

8.2.2 Casos particulares

Los principios del estimador de Horvitz-Thompson se establecen para llegar a una aproximación y estimación de la varianza del estimador. Para los siguientes diseños de muestreo se tienen las siguientes propiedades

Muestreo aleatorio simple

Para este diseño de muestreo en particular las probabilidades de inclusión de primer orden están dadas por $\pi_k = \frac{n}{N}$. Los estimadores de Horvitz-Thompson para las dos características de interés están dados por $\hat{t}_{y,\pi} = N\bar{y}_S$ y $\hat{t}_{z,\pi} = N\bar{z}_S$. Por lo tanto se tiene el siguiente resultado.

Resultado 8.2.2. *Bajo muestreo aleatorio simple, el estimador de la razón poblacional B, su varianza y su varianza estimada están dados por*

$$\hat{B} = \frac{\bar{y}_S}{\bar{z}_S} \tag{8.2.12}$$

$$AVar_{MAS}(\hat{B}) = \frac{N^2}{n}\left(1-\frac{n}{N}\right)S^2_{EU} \quad (8.2.13)$$

$$\widehat{Var}_{MAS}(\hat{B}) = \frac{N^2}{n}\left(1-\frac{n}{N}\right)S^2_{es} \quad (8.2.14)$$

respectivamente, con S^2_{EU} y S^2_{es} el estimador de la varianza de los valores de la variable linealizada E y su aproximación e en el universo U y en la muestra s. Recuerde que $E_k = \frac{1}{t_x}(y_k - Bz_k)$ y $e_k = \frac{1}{\hat{t}_{z,\pi}}(y_k - \hat{B}z_k)$.

Muestreo aleatorio simple en dos etapas

Para este diseño de muestreo los estimadores de Horvitz-Thompson para las dos características de interés están dados por $\hat{t}_{y,\pi} = (N_I/n_I)\sum_{i\in S_I} N_i\bar{y}_{S_i}$ y $\hat{t}_{z,\pi} = (N_I/n_I)\sum_{i\in S_I} N_i\bar{z}_{S_i}$. Se tiene el siguiente resultado.

Resultado 8.2.3. *Bajo muestreo aleatorio simple, el estimador de la razón poblacional B, su varianza y su varianza estimada están dados por*

$$\hat{B} = \frac{\sum_{i\in S_I} N_i\bar{y}_{S_i}}{\sum_{i\in S_I} N_i\bar{z}_{S_i}} \quad (8.2.15)$$

$$AVar_{MM}(\hat{B}) = \frac{N_I^2}{n_I}\left(1-\frac{n_I}{N_I}\right)S^2_{t_E U_I} + \frac{N_I}{n_I}\sum_{i\in U_I}\frac{N_i^2}{n_i}\left(1-\frac{n_i}{N_i}\right)S^2_{y_{E_i}} \quad (8.2.16)$$

$$\widehat{Var}_{MM}(\hat{B}) = \frac{N_I^2}{n_I}\left(1-\frac{n_I}{N_I}\right)S^2_{\hat{t}_e S_I} + \frac{N_I}{n_I}\sum_{i\in S_I}\frac{N_i^2}{n_i}\left(1-\frac{n_i}{N_i}\right)S^2_{e_{S_i}} \quad (8.2.17)$$

respectivamente. Donde $S^2_{t_E U_I}$ es la varianza poblacional de los totales t_{Ei} $i \in U_I$ de todas y cada una de las unidades primarias de muestreo y $S^2_{E_{U_i}}$ es la varianza poblacional entre los valores de la variable E que toman los elementos dentro de cada unidad primaria de muestreo. El razonamiento es similar con las cantidades $S^2_{\hat{t}_e s_I}$ y $S^2_{y_{e_i}}$.

Diseños de muestreo con probabilidad proporcional

Siguiendo con la regla de oro de la estimación de totales, tanto en estrategias que utilicen diseños de muestreos sin reemplazo como Poisson o πPT junto con el estimador de Horvitz-Thompson y en diseños de muestreo con reemplazo junto con el estimador de Hansen-Hurwitz, era conveniente que el marco de muestreo adjuntara información auxiliar de tipo continuo para poder construir las probabilidades de inclusión o de selección según el caso.

Por supuesto, en este contexto particular de estimación de razones, el marco de muestreo debe ser aún más generoso tanto así que permita la inclusión de información auxiliar continua que deberá estar correlacionada **no** con las características de interés que intervienen en la razón **sino** con la variable linealizada E. De esta forma, si la variable correlacionada con E es E^*, entonces las probabilidades óptimas de selección estarían dadas por

$$p_k = \frac{E_k^*}{t_{E^*}} \tag{8.2.18}$$

Un razonamiento similar se hace con los diseños de tamaño fijo que utilizan probabilidades proporcionales.

8.2.3 Estimación de un promedio

Uno de los motivos por los cuales se utiliza el estimador $\hat{B}$ es el desconocimiento del total poblacional N en la estimación de la media poblacional $\bar{y}_U$. Incluso si N es conocido, es preferible ignorarlo como lo demuestra el siguiente ejemplo (Lohr 2000). Suponga que por alguna circunstancia, un extraterrestre desea estimar el número promedio de patas que tiene un perro en una ciudad. La ciudad está dividida en dos áreas geográficas, la zona norte y la zona sur. Para llevar a cabo la estimación, él planea un diseño de muestreo en dos etapas así: De las $N_I = 2$ zonas geográficas de la ciudad va a seleccionar una muestra aleatoria simple de $n_I = 1$ unidades primarias de muestreo. Se sabe que en el norte hay $N_1 = 30$ perros y en el sur hay $N_2 = 10$ perros. Sea cual sea la unidad primaria seleccionada, se seleccionará una sub-muestra aleatoria simple de $n_i = 2$ perros $i = 1, 2$ y se realizará la medición del total de patas en cada perro incluido en la muestra.

Suponga que se ha seleccionado la zona norte. Curiosamente, en esta zona cada uno de los perros tiene igual número de patas, 4. El estimador de Horvitz-Thompson del total de patas en la zona norte está dado por $\hat{t}_{1y,\pi} = \frac{30}{2}8 = 120$. Luego un estimador insesgado del número total de patas en la ciudad está dado por $\hat{t}_{y,\pi} = \frac{2}{1}120 = 240$. Al dividir esta estimación por el número total de perros en la ciudad encontramos la sorpresa de que la estimación de este promedio es 6.

$$\hat{\bar{y}}_{U,\pi} = \frac{\hat{t}_{y,\pi}}{N} = \frac{240}{40} = 6$$

¡¡¡6 patas!!!. Si la muestra del extraterrestre hubiera consistido en la zona sur, el estimador de Horvitz-Thompson del total de patas en la zona sur estría dado por $\hat{t}_{2y,\pi} = \frac{10}{2}8 = 40$. El estimador insesgado del número total de patas en la ciudad estaría dado por $\hat{t}_{y,\pi} = \frac{2}{1}40 = 80$. Al dividir esta estimación por el número total de perros en la ciudad encontramos que la estimación de este promedio es

$$\hat{\bar{y}}_{U,\pi} = \frac{\hat{t}_{y,\pi}}{N} = \frac{80}{40} = 2$$

Sin embargo, a pesar de estos resultados el estimador es efectivamente insesgado porque la esperanza corresponde al parámetro poblacional pues $(2 + 6)/2 = 4$. Seguramente, el extraterrestre no hizo uso de la mejor estrategia de muestreo. No por la escogencia del diseño, que induce probabilidades de inclusión constantes como lo son los valores de la características de interés, sino por el contrario, debido a la escogencia del estimador. Si el estimador utilizado hubiese sido $\hat{B} = \widetilde{y}_S$, definido en (2.2.15), se encontraría que la estimación sería

$$\widetilde{y}_S = \frac{\hat{t}_{y,\pi}}{\hat{N}} = \frac{240}{60} = 4$$

Al seleccionar la zona norte, debido a que $\hat{N} = \frac{2}{1}30 = 60$. Ahora, si hubiese sido seleccionada la zona sur, tendríamos que $\hat{N} = \frac{2}{1}10 = 20$ y por consiguiente

$$\widetilde{y}_S = \frac{\hat{t}_{y,\pi}}{\hat{N}} = \frac{80}{20} = 4$$

Nótese que, para este caso particular, el estimador $\widetilde{y}_S$ es insesgado y de varianza nula. El siguiente resultado amplia las propiedades de este estimador que en la literatura clásica es llamado **promedio muestral ponderado**.

Resultado 8.2.4. *Un estimador del promedio poblacional $\bar{y}_U$, definido como una razón, su varianza y su varianza estimada están dados por*

$$\widetilde{y}_S = \frac{\hat{t}_{y,\pi}}{\hat{N}_\pi} = \sum_S \frac{y_k}{\pi_k} \Big/ \sum_S \frac{1}{\pi_k}. \tag{8.2.19}$$

$$AVar(\widetilde{y}_S) = \frac{1}{N^2} \sum\sum_U \Delta_{kl} \left(\frac{y_k - \bar{y}_U}{\pi_k}\right) \left(\frac{y_l - \bar{y}_U}{\pi_l}\right) \tag{8.2.20}$$

$$\widehat{Var}(\widetilde{y}_S) = \frac{1}{\hat{N}^2} \sum\sum_S \frac{\Delta_{kl}}{\pi_{kl}} \left(\frac{y_k - \widetilde{y}_S}{\pi_k}\right) \left(\frac{y_l - \widetilde{y}_S}{\pi_l}\right) \tag{8.2.21}$$

respectivamente.

Este estimador coincide con el estimador clásico $\bar{y}_S$ en diseños de muestreo como el aleatorio simple o el aleatorio estratificado.

Estimación de un promedio en un dominio

Es la regla, más que la excepción, que el tamaño absoluto N_d de un dominio en estudio sea desconocido. En la sección 3.2.4. se dieron las bases para la estimación del promedio de la característica de interés en un dominio cuando se usaba muestreo aleatorio simple, en esta sección se darán las pautas necesarias para realizar esta estimación bajo cualquier diseño de muestreo y con el desconocimiento de N_d. Siguiendo con la notación de la sección 3.2.4., en donde se definió la función indicatriz del dominio U_d dada por (3.2.22) y se construyó la variable y_{dk}, se tienen los siguientes resultados para la estimación de N_d y para la estimación del total de la característica de interés t_{yd} en el dominio U_d.

Resultado 8.2.5. *Bajo cualquier diseño de muestreo, el estimador de Horvitz-Thompson para el tamaño absoluto de un dominio N_d, su varianza y su varianza estimada están dados por*

$$\hat{N}_{d,\pi} = \sum_S \frac{z_{dk}}{\pi_k} \tag{8.2.22}$$

$$Var(\hat{N}_{d,\pi}) = \sum\sum_U \Delta_{kl} \frac{z_{dk}}{\pi_k} \frac{z_{dl}}{\pi_l} \tag{8.2.23}$$

$$\widehat{Var}(\hat{N}_{d,\pi}) = \sum\sum_S \frac{\Delta_{kl}}{\pi_{kl}} \frac{z_{dk}}{\pi_k} \frac{z_{dl}}{\pi_l} \tag{8.2.24}$$

respectivamente.

Resultado 8.2.6. *Bajo cualquier diseño de muestreo, el estimador de Horvitz-Thompson para el total de la característica de interés t_{yd} en el dominio U_d, su varianza y su varianza estimada están dados por*

$$\hat{t}_{yd,\pi} = \sum_S \frac{y_{dk}}{\pi_k} \tag{8.2.25}$$

$$Var(\hat{t}_{yd,\pi}) = \sum\sum_U \Delta_{kl} \frac{y_{dk}}{\pi_k} \frac{y_{dl}}{\pi_l} \tag{8.2.26}$$

$$\widehat{Var}(\hat{t}_{yd,\pi}) = \sum\sum_S \frac{\Delta_{kl}}{\pi_{kl}} \frac{y_{dk}}{\pi_k} \frac{y_{dl}}{\pi_l} \tag{8.2.27}$$

respectivamente.

Una vez que los anteriores parámetros son estimados y siguiendo la expresión (3.2.23) para el promedio de un dominio, procedemos a estimarlo mediante el siguiente resultado.

Resultado 8.2.7. *Un estimador del promedio de un dominio $\bar{y}_{U_d}$, definido como una razón, su varianza y su varianza estimada están dados por*

$$\widetilde{y}_S = \frac{\hat{t}_{y,\pi}}{\hat{N}_\pi} = \sum_S \frac{y_{dk}}{\pi_k} \Big/ \sum_S \frac{z_{dk}}{\pi_k}. \tag{8.2.28}$$

$$AVar(\widetilde{y}_S) = \frac{1}{N_d^2} \sum\sum_U \Delta_{kl} \left(\frac{y_{dk} - \bar{y}_{U_d}}{\pi_k}\right) \left(\frac{y_{dl} - \bar{y}_{U_d}}{\pi_l}\right) \tag{8.2.29}$$

$$\widehat{Var}(\widetilde{y}_S) = \frac{1}{\hat{N}_d^2} \sum\sum_S \frac{\Delta_{kl}}{\pi_{kl}} \left(\frac{y_{dk} - \widetilde{y}_{S_d}}{\pi_k}\right) \left(\frac{y_l - \widetilde{y}_{S_d}}{\pi_l}\right) \tag{8.2.30}$$

respectivamente.

En el caso específico de muestreo aleatorio simple tenemos que la expresión del estimador alternativo del promedio del dominio dada por (3.2.26) coincide con los anteriores resultados.

Ejemplo 8.2.1. Suponga que para la población de ejemplo U se tiene el conocimiento de cada valor de las características de interés `x` e `y`. De tal forma que la razón poblacional entre las dos es 0.7 como lo muestra la siguiente salida.

```
> y <- c(32,34,46,89,35)
> x <- c(52,60,75,100,50)
> B <- sum(y)/sum(x)
> B
[1] 0.7002967
```

Con una muestra aleatoria simple de $n = 2$, realice el cálculo léxico-gráfico del estimador de la razón $\hat{B}$. Repita el ejercicio con una muestra de $n = 4$ y, por último, con una enumeración completa o censo. Concluya que este estimador es consistente.

8.2.4 Marco y Lucy

Siguiendo con el estudio del sector industrial y con base en las anteriores investigaciones, el gobierno quiere estimar la razón entre el ingreso total del sector industrial con respecto al número de trabajadores del mismo. El anterior es un índice de productividad del sector y describe cuánta ganancia le aporta un sólo empleado al sector. Para el gobierno este índice es importante pues con él se construyen políticas de distribución y apoyo financiero entre los sectores económicos del país.

En los capítulos anteriores hemos aprendido cómo sacar muestras y realizar el proceso de estimación para las estrategias propuestas. En este capítulo vamos a hacer uso de las funciones ya establecidas en el paquete `TeachingSampling` para calcular las estimaciones y estimar las respectivas varianzas. Suponga que se utilizó un diseño de muestreo aleatorio simple y que la muestra seleccionada está dada en la respectiva sección de Marco y Lucy en el segundo capítulo de este texto. Con ayuda de las funciones `S.SI` y `E.SI` del paquete `TeachingSampling`[2] se realiza la selección de la muestra y la estimación de los totales, respectivamente. Después de seleccionar la muestra, procedemos a estimar el total poblacional con la función pertinente. Recuérdese que la salida de la función de estimación es de la siguiente forma

```
> E.SI(N,  n,  característica)

Total estimado              Posición 1
Varianza                    Posición 2
coeficiente de variación    Posición 3
```

Una vez ajustados los parámetros de la función se ingresan los valores de la característica de interés y el resultado de la función es una matriz de estimaciones. En la *Posición 1* encontramos la estimación del total, en la *Posición 2* encontramos la varianza estimada y en la *Posición 3* encontramos el coeficiente de variación estimado. Para tener acceso a cada uno de estos datos de manera independiente es necesario indexar la función, de esta manera si se quiere tener solamente la estimación del total poblacional de la característica ingreso es necesario escribir el siguiente comando: `E.SI(N,n,Income)[1]`.

En donde el índice `[1]` implica el primer elemento de la función. Para lograr la estimación de la razón entre las características Ingreso y Empleados debemos estimar sus respectivos totales con ayuda de la función `E.SI` y realizar el cociente entre ellos.

```
> ty  <-  E.SI(N, n, Income)[1]
> tz  <-  E.SI(N, n, Employees)[1]
> B   <-  ty/tz
> B
[1] 6.920102
```

[2]Por supuesto que el diseño de muestreo puede variar. Si se hubiese usado un diseño aleatorio en dos etapas las funciones que se deberían utilizar serían `S.SI` para seleccionar la muestra y `E.2SI` para realizar las estimaciones.

Aunque se dispone de la estimación debemos realizar la estimación de la aproximación de la varianza. Para este propósito creamos las variables e_k $k \in S$ e introducimos sus valores en la función E.SI para llegar a la estimación de la varianza. Como se mencionó anteriormente, este valor de la estimación de la varianza se encuentra indexado en la segunda posición de la función.

```
> ek   <- (1/tz)*(Income-B*Employees)
> AVar <-  E.SI(N, n, ek)[2]
> cve  <-  100*sqrt(AVar)/B
> cve
[1] 2.261207
```

El resultado de la estimación se presenta en la tabla 8.1. Nótese que el valor estimado se encuentra muy cerca del parámetro de interés.

Tabla 8.1: *Muestreo aleatorio simple: estimación de la razón de interés.*

Variable	Razón poblacional	Razón estimada	cve %	Desv. %
Razón	6.81	6.92	2.26	1.57

Por tanto se estima que cada empleado aportó réditos en el sector industrial hasta por un monto de 6.92 millones de dólares en el último año fiscal. Resultaría interesante saber si esta razón es constante para cada nivel del sector o si se presentan diferencias en la razón para cada estrato. Este tema será tratado en el próximo capítulo.

Teorema del límite central

Al meditar en la confiabilidad y precisión del estimador de la razón, surge la siguiente pregunta: ¿es aplicable el uso del teorema del límite central en la estimación por razones?

Siguiendo con los resultados empíricos, en esta sección se realiza una simulación de Monte Carlo, de tamaño 2000, con las variables Ingreso y Empleados. Para cada simulación, se selecciona una muestra y se estima la razón pertinente. El resultado de la simulación es un conjunto de 2000 estimaciones que se plasmaron en histogramas. El ejercicio se realizó para tamaños de muestra 2, 5, 20, 50, 200 y 1000. El resultado gráfico de la simulación se muestra en la siguiente figura.

Para las primeras simulaciones, en donde el tamaño de muestra es pequeño, se nota que la distribución de la razón es sesgada a la derecha y, a medida que el tamaño de muestra crece, la distribución se torna simétrica con respecto al verdadero valor. Por lo anterior, empíricamente y para este ejemplo en particular, se ha probado que la razón entre estas dos características converge a una distribución normal a medida que el tamaño de muestra se incrementa.

8.3 Estimación de una mediana

Una medida de tendencia central comúnmente utilizada es la mediana. Esta medida de centralidad, a diferencia del promedio poblacional, no es fácilmente influenciada

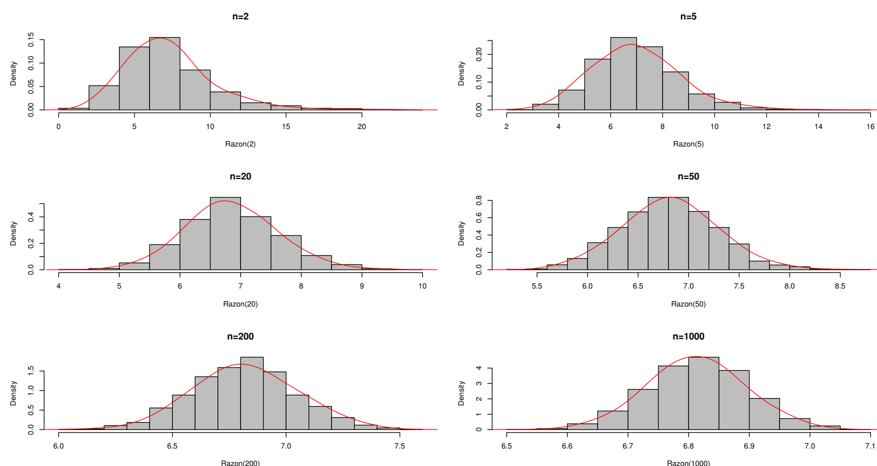


Figura 8.1: *Distribución de muestreo de la razón estimada.*

por datos atípicos cuando el tamaño poblacional es pequeño y, por esto, se conoce como una medida robusta. La mediana es el valor M que divide la población en dos mitades. Por tanto, la mitad de los valores de la característica de interés estará por encima de M y la otra mitad estará por debajo de M. La construcción de esta y otras estimaciones tiene como base la función de distribución poblacional $F(\cdot)$.

Definición 8.3.1. *Para cualquier valor y, la función de distribución poblacional $F(y)$ es la proporción de elementos en la población para los cuales se cumple que $y_k \leq y$. Esta función creciente puede escribirse como*

$$F(y) = \frac{\#A_y}{N} \tag{8.3.1}$$

con A_y dado por

$$A_y = \{k \mid y_k \leq y, k \in U\} \tag{8.3.2}$$

De la anterior definición resulta claro que cualquier percentil[3] Q_q con $0 \leq q \leq 1$ se puede escribir en función de $F(\cdot)$. De esta manera, se tiene que

$$Q_q = F^{-1}(q) \tag{8.3.3}$$

En particular la mediana puede escribirse como $M = Q_{0.5} = F^{-1}(0.5)$. Cuando se ha realizado un diseño de muestreo y la información de la muestra seleccionada es registrada, el procedimiento genérico para la estimación de cualquier percentil sugerido en Särndal, Swensson & Wretman (1992, p. 197) consta de los siguientes pasos:

1. Obtener la función de distribución estimada con los datos de la característica de interés $\widehat{F}(y)$

[3]Valor poblacional para el cual el q % de los valores de la característica de interés en la población cumple que $y_k \leq y$.

2. Estimar el percentil mediante $\widehat{F}^{-1}(q)$. En particular la estimación de la mediana estaría dada por $\widehat{F}^{-1}(0.5)$.

Como lo indican los siguiente resultados no se involucran nuevos principios de estimación en el paso 1 del anterior numeral. El procedimiento para estimar la función de distribución puede verse como la estimación de la media poblacional de la variable z_y que para el k-ésimo elemento de la población está definida como

$$z_{yk} = \begin{cases} 1 & \text{si } y_k \leq y \\ 0 & \text{en otro caso} \end{cases} \tag{8.3.4}$$

Resultado 8.3.1. *La función de distribución poblacional puede escribirse como una función de totales, específicamente como un promedio poblacional y está dada por*

$$\bar{z}_{yU} = \frac{t_{z_y}}{N} = \frac{1}{N}\sum_U z_{yk} = F(y) \tag{8.3.5}$$

Resultado 8.3.2. *Un estimador de la mediana poblacional M está dado por $\hat{M}$*

$$\hat{M} = \widehat{F}^{-1}(0.5), \tag{8.3.6}$$

donde $\widehat{F}^{-1}$ es la función inversa de $\widehat{F}(y)$ dada por

$$\widehat{F}(e) = \frac{\hat{t}_{z_y,\pi}}{\hat{N}} \tag{8.3.7}$$

$$= \sum_S \frac{z_{yk}}{\pi_k}\left(\sum_S \frac{1}{\pi_k}\right)^{-1} \tag{8.3.8}$$

Esta forma de estimación de la mediana arroja los mismos resultados que la estimación de una mediana ponderada[4] por los factores de expansión dados por $1/\pi_k$ $k \in S$. Con este razonamiento concluimos que para los diseños de muestreo que inducen probabilidades de inclusión iguales para cada elemento de la población la estimación de la mediana corresponderá a la mediana de los valores de la característica de interés en la muestra.

Por tanto, si los valores de la característica de interés en la muestra realizada son $\{1, 2, 3\}$ y cada elemento del anterior conjunto está ponderado por su respectivo factor de expansión dado por $\{4, 1, 1\}$, entonces la mediana estimada coincide con la mediana ponderada[5] que es igual a la mediana del siguiente conjunto $\{\underbrace{1, 1, 1, 1}_{4}, \underbrace{2}_{1}, \underbrace{3}_{1}\}$, es decir la mediana es uno.

Ejemplo 8.3.1. Para la población de ejemplo U la mediana poblacional es 35 como lo muestra la siguiente salida.

[4]Draper (1998) afirma que para calcular una mediana ponderada se deben ordenar las observaciones de la menor a la mayor llevando sus pesos a lo largo del ordenamiento. Después es necesario encontrar la suma Σ total de los pesos y añadirlos desde arriba hasta abajo hasta que se encuentre $\Sigma/2$.

[5]Este procedimiento alternativo es computacionalmente mucho más sencillo.

```
> y <- c(32,34,46,89,35)
> median(y)
[1] 35
```

Si el vector de probabilidades de inclusión, inducido por un diseño $p(\cdot)$ de tamaño de muestra fijo e igual $n = 4$, y los factores de expansión están dados por

```
> pik <- c(1,0.5,1,1,0.5)
> fk  <-  1/pik
> fk
[1] 1 2 1 1 2
```

Una posible muestra perteneciente al soporte Q de este diseño de muestreo es

$$s_1\text{=}\{\textbf{Yves, Ken, Erik, Sharon}\}$$

Por tanto la estimación de la mediana para los datos de esta muestra particular será 34 puesto que

```
> w <- c(32,34,34,46,89)
> median(w)
[1] 34
```

¿Cuántas posible muestras tienen probabilidad no nula? Especifique el soporte Q y mediante un cálculo léxico-gráfico concluya acerca del sesgo y de la consistencia del estimador $\hat{M}$.

8.3.1 Marco y Lucy

El gobierno, en su intención de realizar un acercamiento al comportamiento central de las características de interés planeó la investigación de la sección 4.2.4. en donde se planeó un diseño de muestreo con probabilidad proporcional de selección PPT con un tamaño de muestra $m = 400$. En esta ocasión se usó el conocimiento de la característica de interés `Income` para crear las probabilidades de selección de los elementos. Los resultados de la estimación de los totales son verdaderamente cercanos al parámetro de interés por la gran correlación de las probabilidades con las características de interés.

Sin embargo, los investigadores asociados con este proyecto descubren que el comportamiento estructural de la información auxiliar continua Ingreso está influenciado por puntos extremos como se puede ver en la siguiente figura. Por otra parte, se sabe que la correlación entre las características de interés y la información auxiliar es grande, y se supone que el comportamiento estructural de éstas también debe ser muy disperso. Por tanto como medida de centralidad se ha tomado la decisión de trabajar con la mediana porque es una medida robusta.

Una vez que se ha tomado la muestra, siguiendo los pasos de la sección 4.2.4. y con la ayuda de las funciones `S.PPS` y `E.PPS` se utiliza la función `E.Quantile` del paquete `TeachingSampling` para estimar la mediana con la información recolectada en la muestra.

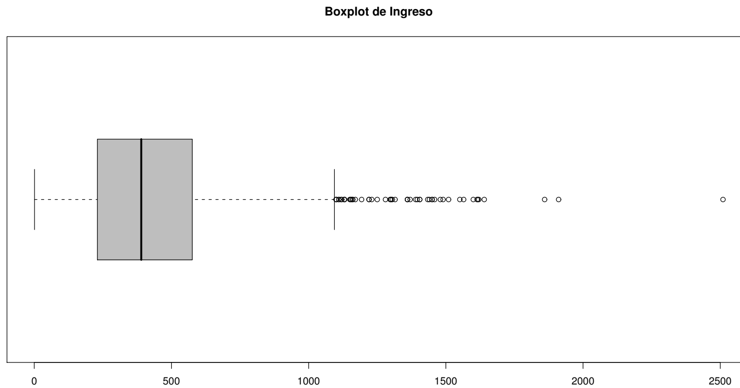


Figura 8.2: *Dispersión de la información auxiliar continua: Ingreso.*

```
> data(Lucy)
> attach(Lucy)
> m <- 400
> sam <- S.PPS(m,Income)
> pk <- Income/sum(Income)
> muestra <- Lucy[sam,]
```

La naturaleza de este ejercicio es muy interesante porque se trata de un diseño con reemplazo. Una vez que la muestra es seleccionada es necesario extraer el vector de probabilidades de selección para las empresas seleccionadas en la muestra. La función `E.Quantile` consta de tres parámetros, `y` que, como de costumbre, es el conjunto de datos conteniendo la información recolectada en la muestra para la(s) característica(s) de interés, `per` que es el percentil de interés y toma valores de 0 a 1, en este caso el valor de interés es 0.5 y corresponde a la mediana y por último `pik` que son las probabilidades de inclusión de cada elemento seleccionado en la muestra[6]. Si este argumento se deja vacío, el resultado de la función será el cálculo del percentil correspondiente para los valores de `y` tratando la muestra como si fuera una población.

```
> pk.s <- pk[sam]
> pik <- 1-(1-pk.s)^400
> attach(muestra)
> estima <- data.frame(Income, Employees, Taxes)
> E.Quantile(estima,0.5,pik)
[1] 420  73  12
```

El resultado de la función arroja las siguientes estimaciones:

- Para la información auxiliar ingreso en el último año fiscal, la mediana estimada es 420 millones de dólares.

[6]En este caso de muestreo PPT utilizamos la expresión (2.2.19) para el cálculo de los π_k a partir de los p_k.

- Para la característica de interés número de empleados, la mediana estimada corresponde a 73.
- Para la característica de interés impuestos declarados en el último año fiscal, la mediana estimada corresponde a 12 millones de dólares.

Si esta muestra se hubiese analizado sin tener en cuenta el diseño de muestreo, las estimaciones serían totalmente diferentes y por lo tanto erradas.

8.4 Estimación de coeficientes de regresión

Hemos llegado a la sección más importante y a la que le da el nombre a esta parte: inferencia asistida por modelos poblacionales. Una vez que hallamos dado los fundamentos teóricos y filosóficos que inspiran un modelo en una población finita, podemos acceder a la mejora de todo tipo de estimadores para la mayoría de parámetros de interés. Es fundamental que el lector, revise una y otra vez la información contenida en esta sección hasta lograr una completa comprensión y apasionamiento por el tema. Una vez que el lector comprenda en su totalidad el espíritu de esta sección estará en capacidad, no sólo de ahondar en temas más complejos e interesantes del muestreo y la inferencia en oblaciones finitas, sino de empezar una rigurosa labor investigativa para crear, construir o mejorar los estimadores propuestos en la literatura clásica.

En la inferencia de poblaciones finitas basada en el diseño de muestreo, se hace hincapié en que las propiedades estadísticas de la estrategia utilizada para la estimación de los parámetros de interés debe estar supeditada al diseño de muestreo que ha usado. Es así como en los capítulos anteriores la esperanza y el cálculo de la varianza y la estimación de la varianza se ha hecho suponiendo un diseño de muestreo $p(\cdot)$ teniendo en cuenta que los valores $y_1, y_2 \ldots, y_N$ que puede tomar la característica de interés son considerados como pseudo-parámetros que son fijos y no son susceptibles de cambio alguno.

Cuando se tiene conocimiento de información auxiliar de tipo continuo o categórico en el marco de muestreo, decimos que para cada elemento en la población existe un vector de información auxiliar que toma el valor $\mathbf{x}_k$ para la k-ésima unidad. Si este vector contiene p características auxiliares entonces toma la siguiente forma: $\mathbf{x}_k = (x_{1k}, x_{2k}, \ldots, x_{pk})'$.

Sin embargo, cuando se ha propuesto determinar la relación existente entre la característica de interés y la información auxiliar continua o categórica contenida en el marco de muestreo, es necesario acudir a un modelo probabilístico que requiere otro tipo de supuestos, que si bien hay que tratar con mucho cuidado, no van en contravía con la teoría propuesta hasta el momento.

8.4.1 Fundamentos teóricos

Suponga que existen N variables aleatorias $Y_1, Y_2, \ldots, Y_N$ por un lado y, que existe un vector de variables aleatorias $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N$ y que la relación entre estas variables aleatorias está dada por un modelo de probabilidad ξ[7] de tal forma que

[7] A este modelo se le conoce con el nombre de modelo se super-población entre Y y $\mathbf{X}$.

$$Y_k = \mathbf{X}_k'\boldsymbol{\beta} + \varepsilon_k \tag{8.4.1}$$

Donde cada un de los ε_k $k \in U$ son variables aleatorias independientes e idénticamente distribuidas con media cero y varianza[8] $c_k\sigma^2$. Al vector $\boldsymbol{\beta}$ se le conoce como vector de coeficientes de regresión en el modelo de super-población o **super-parámetro de regresión**. Bajo las variables ε_k se tienen las siguientes propiedades.

Resultado 8.4.1. *La esperanza y varianza de las variables aleatorias Y_k están dadas por*

$$\begin{aligned} E_\xi(Y_k) &= \mathbf{X}_k'\boldsymbol{\beta} \\ Var_\xi(Y_k) &= c_k\sigma^2. \end{aligned} \tag{8.4.2}$$

Prueba. Las propiedades estadísticas conciernen con el modelo ξ propuesto y con ε_k suponiendo que la información auxiliar es fija. De esta forma

$$\begin{aligned} E_\xi(Y_k) &= E_\xi\left(\mathbf{X}_k'\boldsymbol{\beta} + \varepsilon_k\right) \\ &= \mathbf{X}_k'\boldsymbol{\beta} + E_\xi(\varepsilon_k) \\ &= \mathbf{X}_k'\boldsymbol{\beta}. \end{aligned}$$

Por otro lado, se tiene que

$$\begin{aligned} Var_\xi(Y_k) &= Var_\xi\left(\mathbf{X}_k'\boldsymbol{\beta} + \varepsilon_k\right) \\ &= Var_\xi(\varepsilon_k) \\ &= c_k\sigma^2. \end{aligned}$$

Nótese que el sub-índice ξ denota que la inferencia se realiza bajo la función de distribución inducida por el modelo. ■

Bajo este modelo de super-población los valores $y_1, y_2, \ldots, y_N$ para la característica de interés se consideran realizaciones de las variables aleatorias $Y_1, Y_2, \ldots, Y_N$, lo mismo sucede con los valores del vector $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ que se consideran realizaciones de los vectores aleatorios $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N$. El modelo ξ dado por (8.4.1) y por (8.4.2) es muy general y permite toda clase de acepciones. Pero antes de adentrarnos en cada posible modelo de interés es necesario ahondar un poco más dentro de los fundamentos filosóficos del mismo.

Bajo el modelo ξ se supone una relación entre variables aleatorias dada por el vector de coeficientes de regresión $\boldsymbol{\beta}$ y por las variables aleatorias ε_k. Cassel, Särndal & Wretman (1976*a*) afirman que a ξ se le conoce como modelo de super-población porque supone que la población finita U se toma como si hubiese sido seleccionada de un universo aún más grande al que pertenecen todo tipo de valores para Y_k y para $\mathbf{X}_k$. Dado que es imposible para el hombre calcular el valor de $\boldsymbol{\beta}$ porque, de alguna manera, no está condicionado para conocer el estado de la naturaleza del modelo en cuestión, $\boldsymbol{\beta}$ debe ser estimado usando los datos de la población finita $Y_1, Y_2, \ldots, Y_N$ y $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ mediante la realización de un censo.

[8]Las propiedades estadísticas de estas variables aleatorias deben ser consideradas bajo el modelo ξ.

8.4.2 Estimación en la población finita

Cuando se tiene acceso a la información recolectada en el censo; es decir, se tiene el conocimiento de las realizaciones dadas por y_k y $\mathbf{x}_k$ $(k \in U)$, una forma de estimar, aunque no la única, el super-parámetro de regresión $\boldsymbol{\beta}$ es utilizar el método de los mínimos cuadrados, el cual arrojará como resultado un estimado $\mathbf{B}$.

Dentro del rango de posibles valores que el estimador B pueda tomar, el método de mínimos cuadrados asigna a B el valor que minimiza la siguiente función:

$$D = \sum_U \left(\frac{y_k - \mathbf{x}_k'\mathbf{B}}{c_k\sigma^2} \right)^2. \tag{8.4.3}$$

Una vez más, nótese que ni y_k ni $\mathbf{x}_k$ son variables aleatorias, sino que deben ser tratadas como una realización de variables aleatorias. De esta manera se supone que la relación induce un vector de coeficientes de regresión estimados en la población finita U que pueden ser obtenidos al ajustar el hiperplano $y_k = B_1x_{1k} + \ldots + B_px_{pk}$ para los N elementos en la población entera. El siguiente resultado muestra la forma del estimador de mínimos cuadrados. Para la mejor comprensión de los resultado expuestos en esta sección se escribirán algunas expresiones en lenguaje matricial, así el lector estará familiarizado rápidamente con los modelos lineales.

Resultado 8.4.2. *Usando el método de mínimos cuadrados, el estimador de $\boldsymbol{\beta}$ en la población finita U está dado por*

$$\mathbf{B} = (B_1, \ldots, B_p)' = \left(\mathbf{x}\Sigma^{-1}\mathbf{x}'\right)^{-1}\left(\mathbf{x}\Sigma^{-1}\mathbf{y}\right) \tag{8.4.4}$$

$$= \left(\sum_U \frac{\mathbf{x}_k\mathbf{x}_k'}{c_k\sigma^2}\right)^{-1} \sum_U \frac{\mathbf{x}_k y_k}{c_k\sigma^2} \tag{8.4.5}$$

$$= \left(\sum_U \frac{\mathbf{x}_k\mathbf{x}_k'}{c_k}\right)^{-1} \sum_U \frac{\mathbf{x}_k y_k}{c_k} \tag{8.4.6}$$

Donde

$$\mathbf{x} = \begin{pmatrix} x_{11} & \ldots & x_{1N} \\ \vdots & \ddots & \vdots \\ x_{p1} & \ldots & x_{pN} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1 & \ldots & \mathbf{x}_N \end{pmatrix}; \qquad \mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}. \tag{8.4.7}$$

y Σ es una matriz diagonal de tamaño $N \times N$ dada por

$$\Sigma = \begin{pmatrix} c_1\sigma^2 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & c_N\sigma^2 \end{pmatrix} \tag{8.4.8}$$

Prueba. La expresión que se quiere minimizar es (8.4.3) y corresponde a la suma de cuadrados de los errores $\mathbf{E} = \mathbf{y} - \mathbf{x}'\mathbf{B}$ ponderada por $c_k\sigma^2$ y se puede reescribir de la siguiente forma

$$\begin{aligned} D &= \mathbf{E}'\Sigma^{-1}\mathbf{E} \\ &= (\mathbf{y} - \mathbf{x}'\mathbf{B})'\Sigma^{-1}(\mathbf{y} - \mathbf{x}'\mathbf{B}) \\ &= \mathbf{y}'\mathbf{y} - 2\mathbf{B}'\mathbf{x}\Sigma^{-1}\mathbf{y} + \mathbf{B}'\mathbf{x}\Sigma^{-1}\mathbf{x}'\mathbf{B} \end{aligned}$$

Diferenciando con respecto a $\mathbf{B}$ e igualando a cero

$$\frac{\partial D}{\partial \mathbf{B}} = -2\mathbf{x}'\Sigma^{-1}\mathbf{y} + 2\mathbf{x}'\Sigma^{-1}\mathbf{x}\mathbf{B} \equiv 0$$

encontramos la demostración del resultado. ■

Aunque no es el único método, la técnica de mínimos cuadrados sobresale por sus características de estimación, seguramente el lector deberá estar familiarizado con los métodos de regresión aunque para el lector neófito se sugiere el seguimiento de Ravishanker & Dey (2002) para una buena comprensión de la teoría de modelos lineales. Existen otro tipo de enfoques para la estimación de B, como por ejemplo las técnicas de regresión local polinomial (Breidt & Opsomer 2000) o las técnicas robustas no paramétricas (Gutiérrez 2009*b*, Gutiérrez & Breidt 2009). Es fundamental que el lector note que en la fundamentación teórica nunca se hizo supuesto alguno acerca de la función de distribución de las variables aleatorias ε_k y por lo tanto la inferencia sigue estando libre de asunciones acerca de distribuciones teóricas.

8.4.3 Estimación en la muestra

Por supuesto, en la práctica no tenemos acceso a todos los valores de las característica de interés, incluso en muchas ocasiones no tenemos acceso a todos los valores de la información auxiliar para cada elemento en la población finita. Así que es necesario estimar el coeficiente de regresión. Para este fin y siguiendo con los lineamentos de la sección introductoria se expresa B como una función de totales. En efecto, tenemos que:

$$\mathbf{B} = \mathbf{T}^{\mathbf{-1}}\mathbf{t} \tag{8.4.9}$$

donde

$$\mathbf{T} = \sum_U \frac{\mathbf{x}_k\mathbf{x}_k'}{c_k} \tag{8.4.10}$$

y

$$\mathbf{t} = \sum_U \frac{\mathbf{x}_k y_k}{c_k} \tag{8.4.11}$$

Resultado 8.4.3. *Usando los principios de estimación de una función de totales, cuando el método de mínimos cuadrados es usado,* $\mathbf{B}$ *es estimado por*

$$\hat{\mathbf{B}} = \hat{\mathbf{T}}^{-1}\hat{\mathbf{t}} \tag{8.4.12}$$

donde

$$\mathbf{T} = \sum_S \frac{\mathbf{x}_k\mathbf{x}_k'}{\pi_k c_k} \tag{8.4.13}$$

y

$$\mathbf{t} = \sum_S \frac{\mathbf{x}_k y_k}{\pi_k c_k} \tag{8.4.14}$$

Nótese que $\hat{\mathbf{T}}$ y $\hat{\mathbf{t}}$ son estimadores insesgados para $\mathbf{T}$ y $\mathbf{t}$ respectivamente. Sin embargo, $\hat{\mathbf{B}}$ no es insesgado para $\mathbf{B}$.

Aunque el estimador de $\mathbf{B}$ es sesgado, se debe encontrar una expresión para la varianza. Särndal, Swensson & Wretman (1992) muestran que cuando se usa el método de linealización de Taylor, la aproximación de la varianza del estimador (8.4.12) está dada por

$$AV(\hat{\mathbf{B}}) = \left(\sum_U \frac{\mathbf{x}_k\mathbf{x}'_k}{\sigma^2_k}\right)^{-1} \mathbf{V} \left(\sum_U \frac{\mathbf{x}_k\mathbf{x}'_k}{\sigma^2_k}\right)^{-1}, \tag{8.4.15}$$

donde $\mathbf{V}$ es una matriz simétrica de tamaño $p \times p$ cuyas entradas son

$$v_{ij} = \sum\sum_U \Delta_{kl} \left(\frac{x_{ik}E_k}{\pi_k}\right)\left(\frac{x_{jl}E_l}{\pi_l}\right) \tag{8.4.16}$$

y $E_k = y_k - \mathbf{x}'_k\mathbf{B}$. El estimador de la aproximación de la varianza es

$$\widehat{Var}(\hat{\mathbf{B}}) = \left(\sum_s \frac{\mathbf{x}_k\mathbf{x}'_k}{\sigma^2_k\pi_k}\right)^{-1} \hat{\mathbf{V}} \left(\sum_s \frac{\mathbf{x}_k\mathbf{x}'_k}{\sigma^2_k\pi_k}\right)^{-1}, \tag{8.4.17}$$

donde $\mathbf{V}$ es una matriz simétrica de tamaño $p \times p$ cuyas entradas son

$$\hat{v}_{ij} = \sum\sum_s \frac{\Delta_{kl}}{\pi_{kl}} \left(\frac{x_{ik}e_k}{\pi_k}\right)\left(\frac{x_{jl}e_l}{\pi_l}\right) \tag{8.4.18}$$

y $e_k = y_k - \mathbf{x}'_k\hat{\mathbf{B}}$. Note que $i, j = 1, \ldots, p$.

8.4.4 Casos especiales

El modelo lineal general, definido por las expresiones (8.4.1) y (8.4.2), incluye muchos casos especiales de potencial interés en la práctica para el usuario que desea verificar o estimar la relación existente entre la característica de interés y la información auxiliar. Nótese que este modelo general no tiene restricción alguna en cuanto a la naturaleza de la información auxiliar. Es decir, el vector de información auxiliar $\mathbf{x}_k$ puede ser continuo o categórico.

Existen tres conceptos de vital importancia que se relacionan con la interpretación y el ajuste de cualquier modelo en una población finita. Estos son:

- **Nivel del modelo:** especifica la unidad muestral que se utiliza en la formulación del modelo. Se dice que un modelo se ajusta al nivel de los elementos cuando éste está formulado en términos de información auxiliar disponible

para todos los elementos de la población finita U. Un modelo puede ser formulado tanto a nivel de los elementos como a nivel de conglomerados. Para diseños en varias etapas es posible formular una gran cantidad de modelos a diferentes niveles.

- **Tipo de modelo:** este concepto se refiere al ajuste del mejor modelo que logre explicar la relación entre la característica de interés y la información auxiliar. ¿cuántas variables debo incluir en el modelo? ¿qué estructura de varianza debo proponer? ¿debe tener intercepto el modelo?

- **Modelo de grupo:** cuando se sabe que la población finita U puede ser particionada en grupos poblacionales, es posible ajustar un modelo general que ajuste bien en la población finita. Sin embargo, cuando se sabe que esta partición afecta el comportamiento estructural de la característica de interés en cada grupo, es recomendable ajustar un modelo en cada grupo. Así si la población está compuesta por G grupos, se ajustarán G modelos a cada grupo. Nótese que esta partición puede estar dada tanto a nivel de los elementos como al nivel de la población.

Aunque el modelo lineal general aplica para muchos casos y es obligación del usuario estar en la capacidad de proponer el mejor modelo. Como el maestro Bengt Swensson afirmó en una entrevista concedida en 2005:

> [El modelo lineal general] afirma que existe una relación entre la información auxiliar. Para mí, esos son sólo datos que no traen ninguna información por sí mismos. Sin embargo tienen el **potencial de hacerlo**. Si los datos son útiles en la estimación o no, dependerá de la manera en que $\mathbf{x}$ este relacionado con $\mathbf{y}$. Si el conocimiento y experiencia del estadístico (basados en la realización de anteriores encuestas, muestras piloto o en cualquier otra evidencia) le dicen que efectivamente $\mathbf{x}$ tiene una fuerte relación con $\mathbf{y}$, entonces el modelo comienza a tener sentido. Entre más conocimiento se tenga, se ajustará un mejor modelo.

Con respecto al tipo de modelo es común que en la literatura clásica encontremos los siguientes:

- **Modelo de media común:** este modelo supone que la característica de interés tiene la misma relación común para todo elemento en la población y que la estructura de varianza es constante. Así que $p = 1$, $\mathbf{x}_k = 1$ y $c_k = 1$ para todo $k \in U$. La formulación del modelo está dada por

$$Y_k = \beta + \varepsilon_k \tag{8.4.19}$$

Donde cada un de los ε_k $k \in U$ son variables aleatorias independientes e idénticamente distribuidas con media cero y varianza σ^2.

La figura 8.3 muestra el comportamiento de la relación entre la información auxiliar y la característica de interés. Este modelo tiene las siguientes propiedades:

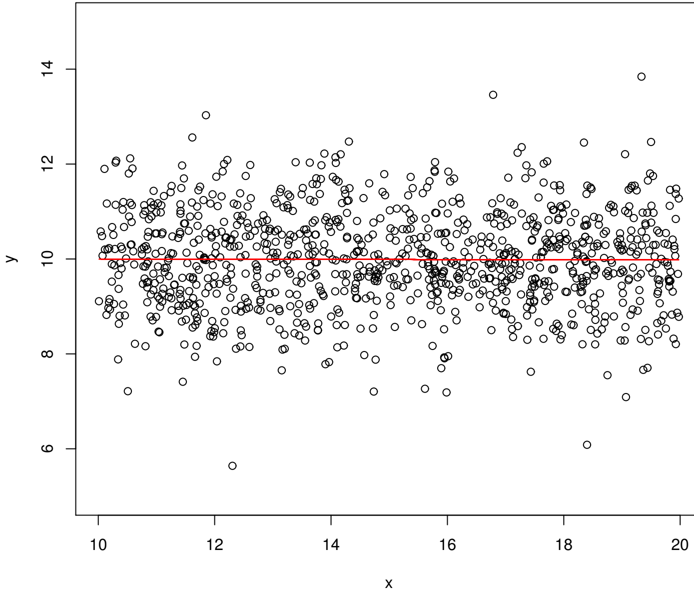


Figura 8.3: *Gráfico de dispersión de un modelo de media común.*

$$\begin{aligned} E_\xi(Y_k) &= \beta \\ Var_\xi(Y_k) &= \sigma^2. \end{aligned} \tag{8.4.20}$$

El estimador del coeficiente de regresión basado en la muestra está dado por

$$\hat{\mathbf{B}} = \left(\sum_S \frac{1}{\pi_k}\right)^{-1} \left(\sum_S \frac{y_k}{\pi_k}\right) = \frac{\hat{t}_{y,\pi}}{\hat{N}_\pi} = \widetilde{y}_S \tag{8.4.21}$$

Luego, bajo este modelo el estimador alternativo del promedio o promedio muestral ponderado es un caso particular del coeficiente de regresión.

- **Modelo de razón:** este modelo supone que la existencia de una sola variable de información auxiliar continua relacionada con la característica de interés y que la estructura de varianza es inversamente proporcional al comportamiento estructural de la información auxiliar. Así que $p = 1$, $\mathbf{x}_k = x_k$ y $c_k = x_k$ para todo $k \in U$. La formulación del modelo está dada por

$$Y_k = X_k'\beta + \varepsilon_k \tag{8.4.22}$$

Donde cada uno de los ε_k $k \in U$ son variables aleatorias independientes e idénticamente distribuidas con media cero y varianza $x_k\sigma^2$.

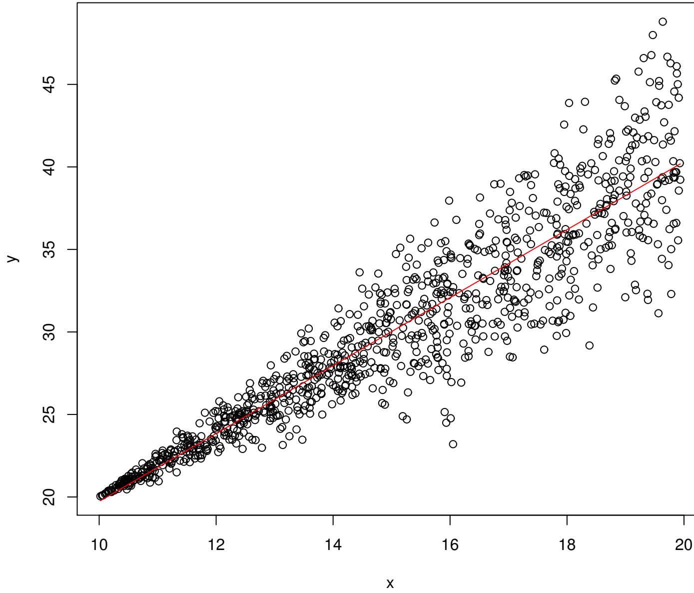


Figura 8.4: *Gráfico de dispersión de un modelo de razón.*

La figura 8.4. muestra el comportamiento de la relación entre la información auxiliar y la característica de interés. Este modelo tiene las siguientes propiedades:

$$
\begin{aligned}
E_\xi(Y_k) &= x'_k\beta \\
Var_\xi(Y_k) &= x_k\sigma^2.
\end{aligned}
\tag{8.4.23}
$$

El estimador del coeficiente de regresión basado en la muestra está dado por

$$
\hat{B} = \left(\sum_S \frac{x_k}{\pi_k}\right)^{-1}\left(\sum_S \frac{y_k}{\pi_k}\right) = \frac{\hat{t}_{y,\pi}}{\hat{t}_{x,\pi}} \tag{8.4.24}
$$

Luego, bajo este modelo el estimador de una razón entre dos características de interés resulta ser un caso particular del coeficiente de regresión.

- **Modelo de regresión simple sin intercepto:** este modelo supone que la existencia de una sola variable de información auxiliar continua relacionada con la característica de interés. Además, supone que la relación debe pasar por el origen del plano cartesiano y que la estructura de varianza es constante. Así que $p = 1$, $\mathbf{x}_k = x_k$ y $c_k = 1$ para todo $k \in U$. La formulación del modelo está dada por

$$
Y_k = X'_k\beta + \varepsilon_k \tag{8.4.25}
$$

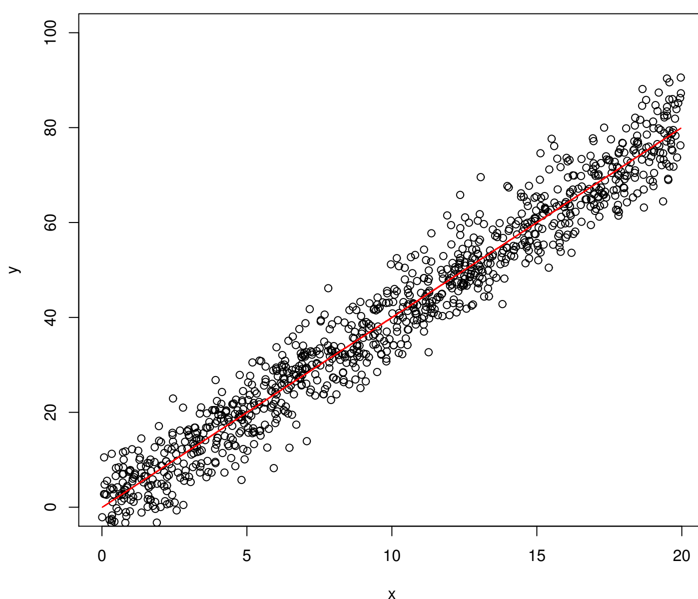


Figura 8.5: *Gráfico de dispersión de un modelo de regresión sin intercepto.*

Donde cada uno de los ε_k $k \in U$ son variables aleatorias independientes e idénticamente distribuidas con media cero y varianza σ^2.

La figura 8.5. muestra el comportamiento de la relación entre la información auxiliar y la característica de interés. Este modelo tiene las siguientes propiedades:

$$\begin{aligned} E_\xi(Y_k) &= x'_k\beta \\ Var_\xi(Y_k) &= \sigma^2. \end{aligned} \tag{8.4.26}$$

El estimador del coeficiente de regresión basado en la muestra está dado por

$$\hat{B} = \left(\sum_S \frac{x_k^2}{\pi_k}\right)^{-1} \left(\sum_S \frac{x_k y_k}{\pi_k}\right) = \frac{\hat{t}_{xy,\pi}}{\hat{t}_{x^2,\pi}} \tag{8.4.27}$$

Es importante resaltar que, al igual que el modelo de razón, éste supone que cuando la característica de interés toma el valor cero, también lo hace la variable de información auxiliar continua.

- **Modelo de regresión simple con intercepto:** este modelo supone que la existencia de dos variables de información auxiliar continuas relacionadas con la característica de interés. Una variable corresponde al vector de unos y la otra corresponde a la información auxiliar continua. Con la inclusión del vector de unos, se supone que la relación no pasa a través del origen. Este

modelo asume que la estructura de varianza es constante. Así que $p = 2$, $\mathbf{x}_k = (1, x_k)'$ y $c_k = 1$ para todo $k \in U$. La formulación del modelo está dada por

$$\begin{aligned} Y_k &= \mathbf{X}_k'\boldsymbol{\beta} + \varepsilon_k \\ Y_k &= \beta_0 + \beta_1 X_k + \varepsilon_k \end{aligned} \tag{8.4.28}$$

Donde cada uno de los ε_k, $k \in U$, son variables aleatorias independientes e idénticamente distribuidas con media cero y varianza σ^2. Para este modelo $\boldsymbol{\beta}' = (\beta_0, \beta_1)$.

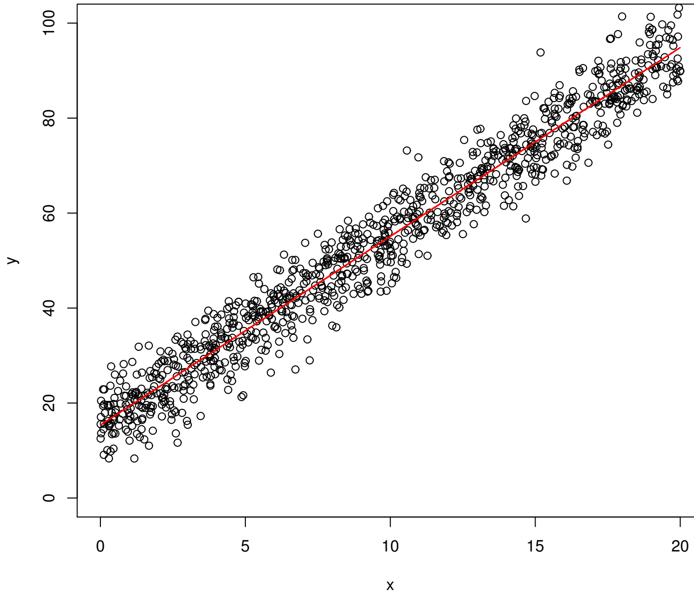


Figura 8.6: *Gráfico de dispersión de un modelo de regresión con intercepto.*

La figura 8.6. muestra el comportamiento de la relación entre la información auxiliar y la característica de interés. Este modelo tiene las siguientes propiedades:

$$\begin{aligned} E_\xi(Y_k) &= \mathbf{x}_k'\boldsymbol{\beta} = \beta_0 + \beta_1 x_k \\ Var_\xi(Y_k) &= \sigma^2. \end{aligned} \tag{8.4.29}$$

El estimador del coeficiente de regresión basado en la muestra está dado por

$$\hat{\mathbf{B}} = \begin{pmatrix} \hat{b}_0 \\ \hat{b}_1 \end{pmatrix} \tag{8.4.30}$$

En donde

$$\hat{b}_1 = \frac{\sum_S \frac{(x_k - \widetilde{x}_S)(y_k - \widetilde{y}_S)}{\pi_k}}{\sum_S \frac{(x_k - \widetilde{x}_S)^2}{\pi_k}} \tag{8.4.31}$$

y

$$\hat{b}_0 = \widetilde{y}_S - \hat{b}_1 \widetilde{x}_S \tag{8.4.32}$$

- **Modelo de media post-estratificada:** este modelo supone la partición en G grupos de la población finita. Así que $U = (U_1, U_2, \ldots, U_G)$. Se asume que la característica de interés está relacionada con G vectores o variables dummy que toman el valor uno si el elemento pertenece al subgrupo U_g $g = 1, \ldots, G$ o cero si el elemento no pertenece al grupo. Así que $p = G$, $\mathbf{x}_k = \mathbf{d}_k = (\underbrace{0, 0, \ldots, 1, \ldots, 0, 0}_{G \texttt{ grupos}})'$ y $c_k = 1$ para todo $k \in U$. La formulación del modelo está dada por

$$Y_k = \mathbf{d}_k' \boldsymbol{\beta} + \varepsilon_k = \beta_g + \varepsilon_k \qquad g = 1, \ldots, G. \tag{8.4.33}$$

Donde $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_g, \ldots, \beta_G)'$ y cada uno de los ε_k $k \in U$ son variables aleatorias independientes e idénticamente distribuidas con media cero y varianza σ_g^2. Nótese que $\mathbf{d}_k = (d_{1k}, \ldots, d_{gk}, \ldots, d_{Gk})'$ con

$$d_{gk} = \begin{cases} 1, & \text{si } k \in U_g \\ 0, & \text{en otro caso.} \end{cases} \tag{8.4.34}$$

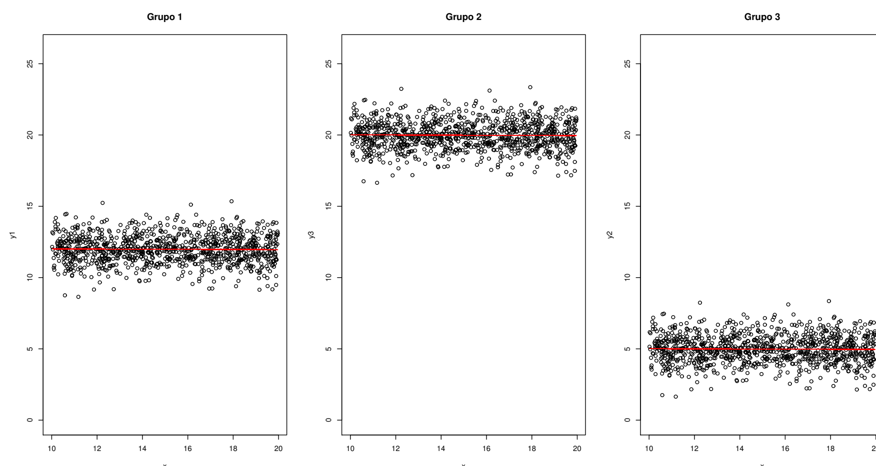


Figura 8.7: *Gráfico de dispersión de un modelo de media post-estratificada.*

La figura 8.7. muestra el comportamiento de la relación entre la información auxiliar y la característica de interés. Este modelo tiene las siguientes propiedades:

$$\begin{aligned} E_\xi(Y_k) &= \mathbf{d}_k'\boldsymbol{\beta} = \beta_g + \varepsilon_k \\ Var_\xi(Y_k) &= \sigma_g^2. \end{aligned} \tag{8.4.35}$$

El estimador del coeficiente de regresión basado en la muestra está dado por

$$\hat{\mathbf{B}} = (\hat{B}_1, \hat{B}_2, \ldots, \hat{B}_G)' \tag{8.4.36}$$

En donde

$$\hat{B}_g = \left(\sum_{S_g} \frac{1}{\pi_k}\right)^{-1} \left(\sum_{S_g} \frac{y_k}{\pi_k}\right) = \frac{\hat{t}_{yU_g,\pi}}{\hat{N}_{U_g,\pi}} = \widetilde{y}_{S_g} \tag{8.4.37}$$

- **Modelo de razón post-estratificada:** este modelo supone la partición en G grupos de la población finita. De tal manera que $U = (U_1, U_2, \ldots, U_G)$. Se asume que es posible definir un modelo de razón en cada uno de los subgrupos U_g $g = 1, \ldots, G$. Así que se considera que la razón entre la característica de interés y la información auxiliar es constante dentro de cada subgrupo pero distinta entre cada subgrupo. Luego, $p = G$, $\mathbf{x}_k = \mathbf{d}_k x_k = \underbrace{(0, 0, \ldots, x_k, \ldots, 0, 0)}_{G \text{ grupos}}'$ y $c_k = x_k$ para todo $k \in U_g$. La formulación del modelo está dada por

$$Y_k = \beta_g X_k + \varepsilon_k \qquad g = 1, \ldots, G. \tag{8.4.38}$$

Donde cada un de los ε_k $k \in U_g$ son variables aleatorias independientes e idénticamente distribuidas con media cero y varianza σ_g^2 para $g = 1, \ldots, G$.

La figura 8.8. muestra el comportamiento de la relación entre la información auxiliar y la característica de interés. Este modelo tiene las siguientes propiedades:

$$\begin{aligned} E_\xi(Y_k) &= \beta_g x_k \\ Var_\xi(Y_k) &= x_k \sigma_g^2. \end{aligned} \tag{8.4.39}$$

El estimador del coeficiente de regresión basado en la muestra está dado por

$$\hat{\mathbf{B}} = (\hat{B}_1, \hat{B}_2, \ldots, \hat{B}_G)' \tag{8.4.40}$$

En donde

$$\hat{\mathbf{B}}_g = \left(\sum_{S_g} \frac{x_k}{\pi_k}\right)^{-1} \left(\sum_{S_g} \frac{y_k}{\pi_k}\right) = \frac{\hat{t}_{yU_g,\pi}}{\hat{t}_{xU_g,\pi}} \tag{8.4.41}$$

Existen más modelos pero los anteriores son los más utilizados en al práctica. La demostración de las anteriores expresiones se deja como ejercicio para el lector.

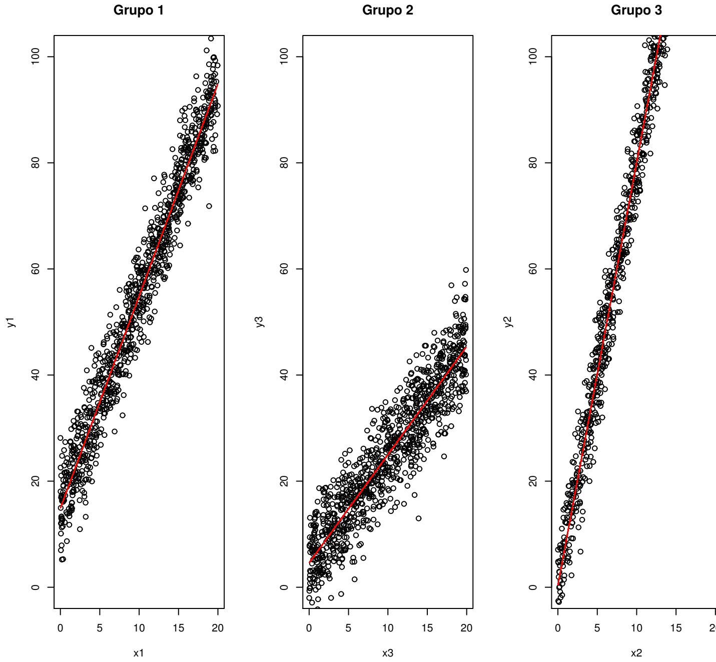


Figura 8.8: *Gráfico de dispersión de un modelo de razón post-estratificada.*

Ejemplo 8.4.1. Retomando nuestra población ejemplo U, suponga que tenemos acceso a los valores de la característica de interés y y de la información auxiliar continua x. Además de esto, se sabe que el modelo que rige la relación entre estas dos está dado por

$$Y_k = \beta_0 + \beta_1 X_k + \varepsilon_k$$

Donde cada uno de los ε_k $k \in U$ son variables aleatorias independientes e idénticamente distribuidas con media cero y varianza constante. Al estimar β_0 y β_1 usando el método de los mínimos cuadrados obtenemos la formulación del modelo en la población finita. Para esto usamos la función `lm` del ambiente computacional de R.

```
> x <- c(32, 34, 46, 89, 35)
> y <- c(52, 60, 75, 100, 50)
> lm(y~x)

(Intercept)             x
     28.505         0.824
```

Lo cual nos lleva a concluir que, en el caso hipotético de tener acceso a todos los datos de la población finita, el modelo estimado sería

$$y_k = 28.505 + 0.824 x_k + E_k$$

Por supuesto, en la práctica no tenemos acceso a la población finita; por tanto, mediante un diseño de muestreo seleccionamos una muestra de tamaño $n = 4$. El diseño de muestreo induce probabilidades de inclusión `pik` para cada uno de los elementos. Suponga que la muestra seleccionada son los primeros cuatro elementos de la población; es decir, **Yves**, **Ken**, **Erik** y por último **Sharon**. Por tanto la información que se ha recolectado después del proceso de medición está guardada en los vectores `x.s` y `y.s`, asimismo lo están las probabilidades de inclusión de los elementos incluidos en la muestra dentro de `pik.s`.

```
> pik <- c(1,0.5,1,1,0.5)
> sam <- c(1,2,3,4)
> x.s <- x[sam]
> y.s <- y[sam]
> pik.s <- pik[sam]
```

Para realizar la estimación de los coeficientes de regresión, se utilizará la función `E.Beta` del paquete `TeachingSampling` que permite la estimación de los coeficientes de regresión bajo cualquier modelo que se proponga con la información recolectada en la muestra. La función `E.Beta` tiene cuatro parámetros los cuales son: `y`, que es el conjunto de datos conteniendo los valores de la(s) característica(s) de interés en la muestra; `x`, que es la matriz de diseño o matriz conteniendo la información auxiliar continua o discreta. Este argumento puede ser un vector, en el caso de una sola variable de información auxiliar, o una matriz, en el caso de múltiple información auxiliar. `pik`, es el vector de probabilidades de inclusión en los elementos incluidos en la muestra. `b0`, que por defecto toma el valor `FALSE` indicando que el modelo fue propuesto sin intercepto. De otra forma, si el modelo propuesto contiene intercepto, `b0` debe tomar el valor `TRUE`. El último argumento de la función es `ck` que hace alusión a la estructura de varianza del modelo, `ck` toma el valor 1 por defecto. Si la estructura de varianza es como el modelo de razón, entonces `ck` deberá ser el mismo vector que se introdujo en el argumento `x`

```
> E.Beta(y.s,x.s,pik.s,b0=TRUE,ck=1)
        [,1]                x
  33.3626289        0.7667526
```

En este caso la estimación, con la información recolectada en la muestra, da como resultado que el intercepto es $\hat{B}_0 = 33.36$ y la pendiente de la recta de regresión es $\hat{B}_1 = 0.77$. La formulación del modelo, en el nivel muestral, estaría dado por:

$$y_k = 33.36 + 0.77x_k + e_k$$

Dado que la estimación de una razón y la media ponderada son casos particulares de la estimación de los coeficientes de regresión, la función `E.Beta` permite fácilmente el cálculo de dichas estimaciones fijando los parámetros de la misma convenientemente.

8.4.5 Marco y Lucy

Es de vital interés para los colaboradores del gobierno conocer la relación entre las características de interés porque con estas relaciones pueden formular modelos econométricos que permitirán ahondar aún más en el comportamiento del sector en el último año fiscal. Si la información poblacional estuviese disponible, y los investigadores estuvieran interesados en formular un modelo distinto para cada las características de interés: número de Empleados y declaración de Impuestos en el último año fiscal con respecto a los Ingresos obtenidos en el mismo.

A continuación, presentamos el razonamiento que nos lleva a escoger el modelo de regresión indicado para cada variable. La información auxiliar continua es la característica Ingreso mientras que las características de interés que tienen relación con esta son Empleados e Impuestos. ¿Tiene sentido ajustar ambos modelos con un intercepto? Piense en el siguiente escenario extremo que se puede presentar... el caso de una empresa que tiene ingresos nulos durante el año pero que aun así sigue funcionando con ayuda del mismo gobierno o con inyección de capital de alguna otra empresa o simplemente con la reserva de capital que la empresa debe guardar. Por lo tanto, si los ingresos son nulos, esto no significa que la empresa tenga cero empleados, entonces es posible que el modelo que se deba ajustar deba tener un intercepto. Por otro lado, si los ingresos son nulos, la declaración de impuestos de la empresa también será nula. Es decir, el modelo que se ajustaría para esta característica de interés no debería contener el parámetro del intercepto.

Entonces, utilizando el método de los mínimos cuadrados estaríamos en capacidad de formular los dos modelos para responder a los objetivos de los investigadores. Ajustamos la regresión utilizando la función `lm`. La estructura de varianza para cada modelo se supone constante.

```
> data(Lucy)
> attach(Lucy)
> y1 <- as.matrix(Employees)
> y2 <- as.matrix(Taxes)
> x <- as.matrix(Income)

> lm(y1~x)                          |          > lm(y2~x-1)
Coefficients:                       |         Coefficients:
(Intercept)            x            |                     x
   29.05839      0.07953            |               0.03648
```

Así que los modelos ajustados en la población finita para las dos características de interés serían

$$Empleados_k = 29.06 + 0.079 \times Ingreso_k + E_k$$

$$Impuestos_k = 0.036 \times Ingreso_k + E_k$$

Por supuesto, los anteriores modelos serían ajustados a la población. En la práctica no tenemos acceso a todos los valores que toman las características de interés,

es por esto que debemos estimar los coeficientes de regresión. Para este fin, el lector deberá retroceder un poco a la sección 7.5. en donde se planeó un diseño de muestreo en dos etapas. La muestra de la primera etapa estaba compuesta por las Zonas **B**,**C**,**D** y **E**. Dentro de cada una de estas zonas geográficas se seleccionó una muestra como lo indica el siguiente código.

```
> muestraI <- c("B","C","D","E")  ## Muestra = "B" "C" "D" "E"
> data(Lucy)
> attach(Lucy)

> Lucy2 <- Lucy[which(Zone==muestraI[1]),]
> Lucy3 <- Lucy[which(Zone==muestraI[2]),]
> Lucy4 <- Lucy[which(Zone==muestraI[3]),]
> Lucy5 <- Lucy[which(Zone==muestraI[4]),]

> N2 <- dim(Marco2)[1];      n2 <- 75
> N3 <- dim(Marco3)[1];      n3 <- 245
> N4 <- dim(Marco4)[1];      n4 <- 50
> N5 <- dim(Marco5)[1];      n5 <- 40
```

Al igual que en la sección 7.5. cuando se selecciona las sub-muestras y la información es recolectada, ésta entra al ambiente computacional de R mediante el conjunto de datos llamado `muestra`. Nótese que el número de filas de `muestra` corresponde al número de empresas encuestadas, es decir, 410. Es necesario tener a mano el vector de probabilidades de inclusión en la muestra cuya longitud también equivale a 410 y está definido por el objeto `pik`.

```
> pik <- rep(c(n2/N2,n3/N3,n4/N4,n5/N5),c(n2,n3,n4,n5))
> muestra <- rbind(muestra2,muestra3,muestra4,muestra5)
> attach(muestra)
```

Para realizar la estimación de los coeficientes de regresión, es necesario utilizar la función `E.Beta` del paquete muestreo. Para el modelo con intercepto de la característica `Employees`, se fijan los parámetros de la función de manera que se ajuste con los preceptos del modelo, note que `b0` toma el valor `TRUE` y que, por la estructura de varianza, `ck` toma el valor `1`. Por otro lado para el modelo sin intercepto de la característica `Taxes`, el valor de `b0` debe ser `FALSE` y al igual que en el modelo anterior, `ck` sigue tomando el valor `1`.

```
> E.Beta(Employees,Income,pik,b0=TRUE,ck=1)
[1,] 25.4329117          [2,]  0.0866556

> E.Beta(Taxes,Income,pik,b0=FALSE,ck=1)
[1,] 0.03739028
```

Así, los modelos estimados en la población finita son

$$Empleados_k = 25.43 + 0.087 \times Ingreso_k + e_k$$

$$Impuestos_k = 0.037 \times Ingreso_k + e_k$$

Esta estimación, a grandes rasgos, indica que, con ingresos nulos, las empresas tienen en promedio a 25 empleados, que cada 11.7 de aumento en ingreso se contrata a un empleado y que en promedio, las empresas pagan una tasa impositiva de 3.7 % al gobierno.

Nótese que si el modelo hubiese sido de razón, entonces la función que se requeriría para la estimación del coeficiente de regresión, que coincide con la estimación de una razón sería:

```
> E.Beta(Taxes,Income,pik,b0=FALSE,ck=Income)
```

8.5 Ejercicios

8.1 Realice el ejercicio lexicográfico del Ejemplo 8.3.1. Ilustre con este ejercicio si el estimador $\hat{M}$ es insesgado o no.

8.2 Con los datos del ejercicio anterior, seleccione una muestra de tamaño $n = 4$. Utilice el resultado 8.3.1 para obtener una estimación de la función de distribución y grafique sus hallazgos.

8.3 Para estimar el total de la característica de interés y de una población de N = 284 elementos, se utilizó un diseño de muestreo Poisson de tamaño de muestra esperado $n(S) = 10$. Las probabilidades de inclusión fueron proporcionales a una característica de información auxiliar x cuyo total poblacional es $t_x = 8182$. El algoritmo de selección arrojó una muestra de tamaño efectivo de 12 elementos, para las cuales se obtuvo la información del ejercicio 4.5. Estime la mediana y las función de distribución para la característica de interés.

8.4 Suponga que los datos del ejercicio anterior fueron obtenidos mediante un diseño de muestreo aleatorio simple. Estime la diferencia de totales $t_y - t_x$ mediante $\hat{t}_{y-x} = \hat{t}_{y,\pi} - \hat{t}_{x,\pi}$. Estime la varianza y calcule el coeficiente de variación estimado.

8.5 Suponga que los datos del ejercicio anterior fueron obtenidos mediante un diseño de muestreo Bernoulli con $\pi = 0.04$.

(a) Estime la razón de totales t_y/t_x mediante $\hat{B} = \hat{t}_{y,\pi}/\hat{t}_{x,\pi}$. Estime la varianza y calcule el coeficiente de variación estimado.

(b) Estime el promedio de la característica de interés utilizando el estimador de Hájek. Estime la varianza y calcule el coeficiente de variación estimado.

(c) Estime el promedio de la característica de información auxiliar utilizando el estimador de Hájek. Estime la varianza y calcule el coeficiente de variación estimado.

8.6 Verifique la expresión de la matriz de varianzas $AV(\hat{\mathbf{B}})$

8.7 En una muestra de municipios, basada en un diseño de muestreo aleatorio simple, se seleccionaron $n = 10$ municipios de $N = 49$. En cada municipio se midieron las siguientes características: el número de habitantes en el municipio (**HAB**), el numero de automoviles en el municipio (**VEH**) y el número de

efectivos militares en el municipio (**MIL**). Además, se sabe que cada municipio se categoriza (**CAT**) en urbano (**CAT**=1) o rural (**CAT**=0). A continuación se muestra la información recolectada de los municipios en la muestra:

HAB	**VEH**	**MIL**	**CAT**
2571	50	415	1
2813	55	462	1
3002	61	513	1
3564	70	577	1
3051	64	532	0
2835	56	463	0
3319	67	551	0
2986	61	512	0
2998	55	471	0
2717	56	462	0

(a) Estime el coeficiente de regresión de **HAB** contra **VEH** para un modelo de media común. Estime la varianza y calcule el coeficiente de variación. Interprete el coeficiente estimado.

(b) Estime el coeficiente de regresión de **HAB** contra **VEH** para un modelo de razón. Estime la varianza y calcule el coeficiente de variación. Interprete el coeficiente estimado.

(c) Estime los coeficientes de regresión de **HAB** contra **MIL** para un modelo de regresión simple con intercepto. Estime la matriz de varianzas y calcule los coeficientes de variación. Interprete los coeficiente estimados.

(d) Estime los coeficientes de regresión de **HAB** contra **CAT** para un modelo de media post-estratificada. Estime la matriz de varianzas y calcule los coeficientes de variación. Interprete los coeficiente estimados.

(e) Estime los coeficientes de regresión de **HAB** contra **MIL** para un modelo de razón post-estratificada mediante **CAT**. Estime la matriz de varianzas y calcule los coeficientes de variación. Interprete los coeficiente estimados.

8.8 Sustente o refute las siguientes afirmaciones

(a) Una función lineal de estimadores insesgados es siempre insesgada para su contraparte poblacional.

(b) Se dice que un estimador es aproximadamente insesgado cuando es sesgado sólo para la parte lineal del desarrollo de Taylor.

(c) En la estimación de una razón poblacional B, se cumple para la variable linealizada $E_k = \frac{1}{t_z}(y_k - Bz_k)$ que $\sum_S E_k = 0$ sin importar el diseño de muestreo utilizado en el planeamiento del estudio.

(d) El estimador $\hat{B} = \frac{\hat{t}_{y,\pi}}{\hat{t}_{z,\pi}}$ es sesgado para $B = \frac{t_y}{t_z}$ sólo si z_k es continua.

(e) En diseños de muestreo de tamaño de muestra aleatorio, el estimador del promedio poblacional $\tilde{y}_S$ es insesgado y de menor varianza en comparación al estimador $\bar{y}_S$.

(f) El método de linealización de Taylor para aproximar la varianza de parámetros complejos y en muestras pequeñas conduce generalmente a la sobreestimación de la varianza real.

(g) El estimador $\hat{\mathbf{B}} = \hat{\mathbf{T}}^{-1}\hat{\mathbf{t}}$ es siempre sesgado para $\mathbf{B} = \mathbf{T}^{-1}\mathbf{t}$ independientemente de la calidad del ajuste.

Capítulo 9

Estimación con información auxiliar

> Si los datos son útiles en la estimación o no, dependerá de la manera en que $\mathbf{x}$ esté relacionado con $\mathbf{y}$. Si el conocimiento y experiencia del estadístico le dicen que efectivamente $\mathbf{x}$ tiene una fuerte relación con $\mathbf{y}$, entonces el modelo comienza a tener sentido. Entre más conocimiento se tenga, se ajustará un mejor modelo.
>
> Bengt Swensson (2005)

Las nociones de la inferencia en poblaciones finitas fueron expresadas hace más de 60 años en muchos libros clásicos como Cochran, Hansen, Hurwitz y Madow, Deming, Muthy, Des Raj y otros. La teoría de muestreo era aplicada desde la perspectiva misma de la selección aleatorizada de posibles muestras en la población finita. Dependiendo de las circunstancias prácticas, la selección se hacía de distintas maneras: muestreo aleatorio simple, muestreo aleatorio estratificado, muestreo de conglomerados, muestreo en dos etapas, etc. El muestreo era considerado como la actividad primaria y la estimación nunca fue considerada como una práctica separada sino como una consecuencia automática. Lo anterior se debía a que cada tipo de diseño de muestreo inducía un estimador cuyas propiedades estadísticas como el insesgamiento y la varianza eran establecidas de antemano con el diseño y así, la varianza era calculable y estimable.

Así que, para la década de los sesenta, muchos creyeron que la investigación en el campo del muestreo y de la inferencia en poblaciones finitas ya estaba muerta porque se deberían inventar nuevas formas de selección de muestras (tarea ardua y difícil), más allá de las que se cubrían en los libros clásicos del muestreo. Aunque el estimador de razón fue considerado en algún detalle por los textos de referencia, la inclusión de varias variables de información auxiliar no se vio como un tópico que prometiera rédito alguno para emprender el camino de la investigación en esa vía. En la década de los setenta, varios autores dieron un viraje en su perspectiva epistemológica de la inferencia en poblaciones finitas. Es así como Basú, Brewer, Godambe y Royall, entre otros, consideraron los modelos estadísticos (en sintonía con la estadística clásica Fisheriana) como los verdaderos fundamentos de la es-

timación e inferencia en poblaciones finitas. Su trabajo se cimentó alrededor de la posibilidad de tener una inferencia que dependiera estrictamente del modelo propuesto y no tuviera nada que ver con el diseño de muestreo utilizado en la recolección de los datos. Como consecuencia, la atención se tornó alrededor de la estimación y se dejo de lado el muestreo por la relación existente o propuesta entre la característica de interés y las variables de información auxiliar.

El camino que tomó la historia del muestreo fue, precisamente, la incorporación de las dos corrientes de pensamiento bajo una sola sombrilla. Así que, fue posible combinar la aleatorización clásica con un percepción más general de la relación de y con x. No hubo necesidad de sacrificar los principios basados en la aleatorización. Así nació la inferencia asistida por modelos pero basada en ala aleatorización (*model assisted design-based inference* por su original en inglés). Este nuevo tipo de inferencia se hizo muy atractiva porque la regresión y los modelos acompañan al estadístico desde sus primeros cursos y van tomando más fuerzas a medida que se avanza en el camino universitario. Así que, este pensamiento ≪asistido por modelos≫ es un matrimonio efectivo y tolerante que permite las ideas de la regresión junto con el paradigma de la aleatorización.

Jan Wrettman (Kott, Sweensson, Särndal & Wretman 2005) opina que el ajuste de un modelo se ha convertido en parte integral de la teoría clásica del muestreo, aunque los principios de la misma deben permanecer intocables porque las propiedades de los estimadores son evaluadas con respecto al mecanismo de probabilidad que genera la muestra y no con respecto a cualquier modelo asumido.

9.1 Introducción

En los capítulos anteriores de este texto, el lector ha sido introducido en los diferentes diseños de muestreo que, dependiendo de la configuración de los valores de la característica de interés, mejoran la eficiencia de los estimadores de Horvitz-Thompson o Hansen-Hurwitz, según sea el caso. En algunas ocasiones, el uso correcto de la información auxiliar en la etapa de diseño hace que la eficiencia de los estimadores mejore dramáticamente. Por ejemplo, si la información auxiliar es de tipo categórico y está bien correlacionada con el comportamiento estructural de la característica de interés, es posible acudir a un diseño de muestreo estratificado. De otra forma, si la información auxiliar disponible en la población es de tipo continuo, podemos utilizar un diseño de muestreo PPT o πPT para mejorar la precisión de las estimaciones. En cualquiera de los casos, es necesario:

1. Conocer los valores de la información auxiliar, ya sea de tipo continua o categórica, para todos los elementos que conforman la población.

2. Tener la certeza de que la característica de interés guarda una estrecha correlación positiva con la información auxiliar.

En este capítulo, el interés está centrado en mejorar la eficiencia de las estimaciones incorporando al estimador la información auxiliar, que puede ser de tipo categórico o continuo, fijando el diseño de muestreo utilizado. En otras palabras, se quiere hacer uso de la información auxiliar en la etapa de estimación. Para este fin es necesario:

1. Contar con la experticia del investigador que ha sabido discernir y escoger el mejor diseño de muestreo para la configuración de los valores de la característica de interés.

2. Saber que la característica de interés está bien relacionada con la información auxiliar. Como se verá más adelante no es necesario el conocimiento estricto de los valores de la información auxiliar en todos los elementos de la población, aunque sí es necesario conocer estos valores para la muestra junto con el total poblacional de la información auxiliar en la población[1].

Por supuesto, los nuevos estimadores, que incorporan información auxiliar, apuntan a la mejora dramática en la eficiencia de las estrategias de estimación de totales poblacionales. Además de esta característica, existen muchas otras que tienen que ver con la consistencia y el insesgamiento. Sin embargo, una característica importante de un estimador construido a partir de la información auxiliar está dada por la siguiente definición.

Definición 9.1.1. *Una estrategia de muestreo se dice* ***representativa*** *con respecto a la información auxiliar* $\mathbf{x}$*, sí y sólo sí*

$$\hat{t}_S(\mathbf{x}) = t_{\mathbf{x}}. \tag{9.1.1}$$

Es decir, si el estimador aplicado a las variables auxiliares reproduce exactamente el total poblacional de las mismas.

La idea detrás del principio de representatividad de la estrategia es que si se tiene el conocimiento de que la característica de interés guarda una estrecha relación lineal con la información auxiliar entonces podemos pensar en que la siguiente igualdad se cumple

$$t_{\mathbf{x}} \approx t_y \tag{9.1.2}$$

y, una consecuencia inmediata de esta propiedad, bajo los anteriores supuestos es que

$$\hat{t}_S(y) \approx t_y \tag{9.1.3}$$

Sin importar el diseño de muestreo utilizado para la selección de la muestra, si el total poblacional de las variables auxiliares, $t_{\mathbf{x}}$, es conocido, se puede utilizar esta información para construir un estimador aún más preciso. En este capítulo se consideran los estimadores lineales de la forma

$$\hat{t}_S(y) = w_0 + \sum_{k\in S} w_k y_k, \tag{9.1.4}$$

En donde los pesos w_k pueden depender del vector de información auxiliar. Es claro que no todos los estimadores lineales cumplen la ecuación de representatividad. Por ejemplo, el estimador de Horvitz-Thompson es insesgado pero no utiliza

[1] Esta información puede ser suministrada por alguna entidad oficial.

información auxiliar por tanto no cumple la ecuación de representatividad para la información auxiliar. Aunque de manera teórica no es difícil mostrar que, utilizando un diseño de muestreo de tamaño de muestra fijo, el estimador de $\hat{t}_{y\pi}$ arroja una estrategia representativa sobre el vector de probabilidades de inclusión $\pi_1, \ldots, \pi_N$.

Si $\hat{t}_{y,\pi}$ y $\hat{t}_{\mathbf{x},\pi}$ son los estimadores de Horvitz-Thompson de y y $\mathbf{x}$ respectivamente, entonces es posible construir nuevos estimadores que, sin importar el diseño de muestreo, arrojen estrategias representativas sobre el vector de información auxiliar $\mathbf{x}$. Bajo estas condiciones la precisión de la estimación queda asegurada mediante la aplicación del siguiente resultado.

Resultado 9.1.1. *Si el estimador $\hat{t}_S(\cdot)$ induce una estrategia representativa sobre el vector de información auxiliar $\mathbf{x}$, tal que (9.1.1) se satisface. Entonces $\hat{t}_S(\mathbf{x})$ estimará el total $t_{\mathbf{x}}$ con varianza nula.*

Prueba. Si (9.1.1) se cumple, entonces

$$Var(\hat{t}_S(\mathbf{x})) = Var(t_{\mathbf{x}}) = 0 \tag{9.1.5}$$

Nótese que el operador $Var(\cdot)$ se calcula sobre todas las posibles muestras del soporte Q inducido por el diseño de muestreo. Es decir, para todas las muestras pertenecientes a Q el estimador $\hat{\mathbf{t}}_S(\mathbf{x})$ reproducirá el total $\mathbf{t}_{\mathbf{x}}$ ■

Este resultado es muy importante porque si es cierto que la característica de interés está relacionada con la información auxiliar, entonces $\hat{t}_S(y)$ tenderá a contar con una varianza muy pequeña.

Ahora es tiempo de discutir sobre la incorporación de la información auxiliar al estimador. ¿Cómo es posible introducir esta información en una expresión matemática que intenta estimar un parámetro? La respuesta es simple y clara: mediante un modelo de super-población ξ.

9.2 Estimador general de regresión

En esta sección se construye un estimador del total poblacional de la característica de interés t_y que mejora dramáticamente en eficiencia al incorporar información auxiliar. La manera en que esta incorporación se realiza es mediante el supuesto de que las variables de información auxiliar están relacionadas con la característica de interés mediante un modelo ξ. Este modelo es un modelo lineal general y le da el nombre al estimador que se propone en este capítulo. Así que si existen N variables aleatorias $Y_1, Y_2, \ldots, Y_N$ y un vector de variables aleatorias $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N$ y la relación entre estas variables aleatorias está dada por un modelo de super-población, de tal forma que:

$$Y_k = \mathbf{X}_k'\boldsymbol{\beta} + \varepsilon_k \tag{9.2.1}$$

Donde cada uno de los ε_k $k \in U$ son variables aleatorias independientes e idénticamente distribuidas con media cero y varianza $c_k\sigma^2$, tales que:

$$\begin{aligned} E_\xi(Y_k) &= \mathbf{X}_k'\boldsymbol{\beta} \\ Var_\xi(Y_k) &= c_k\sigma^2. \end{aligned} \tag{9.2.2}$$

Al considerar este modelo general es posible construir un estimador del total poblacional que conciba esta relación.

9.2.1 Construcción

Sea U el conjunto de elementos en la población finita y S el conjunto de los elementos que conforman la muestra aleatoria. Sean y_k, $k \in S$ y $\mathbf{x}_k$, $k \in U$, los valores de la característica de interés y y el vector de información auxiliar asociados al k-ésimo elemento de la población. Siendo π_k la probabilidad de inclusión de primer orden, se asume que los totales poblacionales de la información auxiliar $\mathbf{t_x} = \sum_{k\in U} \mathbf{x_k}$ son conocidos.

De manera general, se asume que existe una relación entre la variable de interés y la información auxiliar por el modelo de super-población ξ. Es decir,

$$y_k = f(x_{1k}, x_{2k}, \ldots, x_{pk}) + E_k \tag{9.2.3}$$

En particular, bajo ξ existe una relación de tipo lineal entre y_k y $\mathbf{x_k}$. Por tanto, en la población finita se tiene que

$$\begin{aligned} y_k &= \mathbf{x}_k'\mathbf{B} + E_k \\ &= {y_k}^{\mathrm{o}} + E_k \end{aligned}$$

Entonces, el parámetro poblacional que se quiere estimar se puede escribir como

$$t_y = \sum_U ({y_k}^{\mathrm{o}} + y_k - {y_k}^{\mathrm{o}}) \tag{9.2.4}$$

$$= \sum_U \mathbf{x}_k'\mathbf{B} + \sum_U (y_k - {y_k}^{\mathrm{o}}) \tag{9.2.5}$$

$$= \sum_U \mathbf{x}_k'\mathbf{B} + \sum_U E_k \tag{9.2.6}$$

$$= \sum_U {y_k}^{\mathrm{o}} + \sum_U E_k \tag{9.2.7}$$

Como el objetivo es estimar t_y con los datos suministrados en la muestra. Entonces es necesario estimar dos cantidades. La primera es $\mathbf{B}$ que corresponde a un vector de coeficientes de regresión y que puede ser estimado siguiendo los principios del capítulo anterior. La segunda cantidad corresponde al total t_E que puede ser estimado utilizando los principios del estimador de Horvitz-Thompson. De esta manera, se tiene la construcción del estimador general de regresión.

Definición 9.2.1. *El estimador general de regresión está definido por la siguiente expresión*

$$\hat{t}_{y,greg} = \sum_U \mathbf{x}_k'\hat{\mathbf{B}} + \sum_s \frac{y_k - \mathbf{x}_k'\hat{\mathbf{B}}}{\pi_k} \tag{9.2.8}$$

Desarrollando la expresión del estimador general de regresión y factorizando convenientemente, llegamos a que el estimador general de regresión se puede escribir como:

$$\hat{t}_{y,greg} = \sum_U \mathbf{x}_k'\hat{\mathbf{B}} + \sum_s \frac{y_k}{\pi_k} - \sum_s \frac{\mathbf{x}_k'\hat{\mathbf{B}}}{\pi_k} \tag{9.2.9}$$

$$= \hat{t}_{y\pi} + \sum_{j=1}^{J} \hat{B}_j(t_{xj} - \hat{t}_{xj\pi}) \tag{9.2.10}$$

Que matricialmente se deja escribir como:

$$\hat{t}_{y,greg} = \hat{t}_{y\pi} + (\mathbf{t_x} - \hat{\mathbf{t}}_{\mathbf{x}\pi})'\hat{\mathbf{B}} \tag{9.2.11}$$

Como el estimador de $\mathbf{B}$ se halló utilizando la técnica de mínimos cuadrados, entonces

$$\hat{\mathbf{B}} = \hat{\mathbf{T}}^{-1}\hat{\mathbf{t}} \tag{9.2.12}$$

donde

$$\widehat{\mathbf{T}} = \sum_S \frac{\mathbf{x}_k\mathbf{x}_k'}{\pi_k c_k} \tag{9.2.13}$$

y

$$\hat{\mathbf{t}} = \sum_S \frac{\mathbf{x}_k y_k}{\pi_k c_k} \tag{9.2.14}$$

Por tanto, al descomponer $\hat{\mathbf{B}}$[2], el estimador toma la siguiente forma

$$\hat{t}_{y,greg} = \sum_s \frac{y_k}{\pi_k} + (\mathbf{t_x} - \hat{\mathbf{t}}_{\mathbf{x}\pi})'\mathbf{T}^{-1}\sum_s \frac{\mathbf{x}_k y_k}{c_k\pi_k} \tag{9.2.15}$$

$$= \sum_s \left(1 + (\mathbf{t_x} - \hat{\mathbf{t}}_{\mathbf{x}\pi})'\mathbf{T}^{-1}\frac{\mathbf{x}_k}{c_k}\right)\frac{y_k}{\pi_k} \tag{9.2.16}$$

$$= \sum_s g_{ks}\frac{y_k}{\pi_k} \tag{9.2.17}$$

Por lo tanto, se tienen distintas formas de escribir el mismo estimador; las últimas expresiones son particularmente útiles, pues los pesos g_{ks} tienen la propiedad de inducir estrategias representativas sobre cualquier variable del vector auxiliar. Es decir, al aplicar los pesos, sobre la muestra, a una variable de la información auxiliar, el resultado será el total poblacional de dicha variable.

$$\hat{\mathbf{t}}_{\mathbf{x},greg} = \sum_S g_{ks}\frac{\mathbf{x}_k'}{\pi_k} = \mathbf{t_x} \tag{9.2.18}$$

[2]Nótese que $\hat{\mathbf{B}}$ no es un estimador insesgado para $\mathbf{B}$.

Volviendo atrás a la introducción de este capítulo, se puede concluir que el estimador de regresión general es un estimador de tipo lineal con $w_0 = 0$ y $w_k = \frac{g_{ks}}{\pi_k}$. De tal forma que

$$\hat{t}_{y,greg} = \sum_S w_k y_k \tag{9.2.19}$$

$$= \sum_S g_{ks} \frac{y_k}{\pi_k} \tag{9.2.20}$$

con

$$g_{ks} = 1 + (\mathbf{t_x} - \hat{\mathbf{t}}_{\mathbf{x}\pi})'\mathbf{T}^{-1}\frac{\mathbf{x}_k}{c_k} \tag{9.2.21}$$

A los pesos w_k se les conoce con el nombre de **pesos de calibración** y son usados ampliamente en la construcción de estimadores asistidos en modelos de superpoblación. De esta manera, al usar los pesos calibrados el estimador asistido por modelos está dado por

$$\hat{t}_{y,cal} = \sum_{k \in S} w_k y_k. \tag{9.2.22}$$

Nótese que una propiedad de los pesos de calibración es que el estimador de la información auxiliar reproduce exactamente los totales poblacionales de la misma. De esta forma, tenemos que

$$t_{x,cal} = \sum_{k \in S} w_k x_k = t_x. \tag{9.2.23}$$

Resultado 9.2.1. *Para cualquier diseño de muestreo, el estimador $\widehat{t}_{y,greg}$ induce una estrategia representativa sobre el vector de variables auxiliares. Es decir*

$$\widehat{\mathbf{t}}_{\mathbf{x},greg} = \mathbf{t_x} \tag{9.2.24}$$

Prueba. Utilizando la forma matricial del estimador general de regresión dada por la expresión (9.2.11) se tiene que

$$\widehat{\mathbf{t}}_{\mathbf{x},greg} = \hat{\mathbf{t}}_{\mathbf{x}\pi} + (\mathbf{t_x} - \hat{\mathbf{t}}_{\mathbf{x}\pi})'\hat{\mathbf{B}}$$

Sin embargo, $\hat{\mathbf{B}}$ será los coeficiente de regresión, ajustados por mínimos cuadrados, entre la información auxiliar contra ella misma. Por lo tanto, se tratará de una matriz identidad. Esto es claro al desarrollarlo, por tanto

$$\hat{\mathbf{B}} = \widehat{\mathbf{T}}^{-1}\widehat{\mathbf{T}} = \left(\sum_S \frac{\mathbf{x}_k\mathbf{x}_k'}{\pi_k c_k}\right)^{-1}\left(\sum_S \frac{\mathbf{x}_k\mathbf{x}_k'}{\pi_k c_k}\right) = \mathbf{I}_{p\times p}$$

Entonces, el estimador general de regresión del vector de totales de la información auxiliar será

$$\widehat{\mathbf{t}}_{\mathbf{x},greg} = \hat{\mathbf{t}}_{\mathbf{x}\pi} + (\mathbf{t}_{\mathbf{x}} - \hat{\mathbf{t}}_{\mathbf{x}\pi})'\mathbf{I}_{p\times p}$$
$$= \hat{\mathbf{t}}_{\mathbf{x}\pi} + \mathbf{t}_{\mathbf{x}} - \hat{\mathbf{t}}_{\mathbf{x}\pi}$$
$$= \mathbf{t}_{\mathbf{x}}$$

■

Es importante resaltar que la conformación estructural de los pesos de calibración depende de

1. El modelo de superpoblación y sus condicionamientos para la estimación de los parámetros de regresión. Es decir, la forma del modelo per se (con o sin intercepto y la cantidad de variables de información auxiliar) y la estructura de varianza (el valor que toma c_k).
2. El vector de probabilidades de inclusión en la muestra.
3. La muestra realizada. Para cada posible muestra del soporte definido por el diseño de muestreo, existe una configuración distinta de pesos de calibración.

Ejemplo 9.2.1. Retomando nuestra población ejemplo U, suponga que el modelo de super-población ξ es tal que

$$Y_k = \beta_0 + \beta_1 X_k + \varepsilon_k$$

Donde cada uno de los ε_k $k \in U$ son variables aleatorias independientes e idénticamente distribuidas con media cero y estructura de varianza constante. Los valores de la característica de interés y de la información auxiliar continua se muestran a continuación

```
> x <- c(32, 34, 46, 89, 35)
> y <- c(52, 60, 75, 100, 50)
```

Mediante un diseño de muestreo aleatorio simple se selecciona una muestra de tamaño $n = 4$. Por supuesto, este diseño de muestreo induce probabilidades de inclusión `pik` para cada uno de los elementos.

```
> sam <- sample(5,4)
> pik <- rep(4/5,5)
```

Suponga que la muestra realizada está dada por los elementos 1, 2, 3 y 5 de la población, correspondientes a **Yves, Ken, Erik, Leslie**. Los valores de `y`, `x`, y de `pik` para cada uno de los elementos en la muestra están dados por

```
> x.s <- x[sam]
> y.s <- y[sam]
> pik.s <- pik[sam]
```

Con la ayuda de la función Wk del paquete TeachingSampling es posible realizar el cálculo de los pesos de calibración para los elementos seleccionados en la muestra. Esta función tiene cinco argumentos descritos a continuación: x, que es la matriz de información auxiliar conteniendo los valores para cada uno de los elementos de la muestra de la información auxiliar continua o discreta. Este argumento puede ser un vector, en el caso de una sola variable de información auxiliar, o una matriz, en el caso de múltiple información auxiliar. tx, que es el vector de totales poblacionales (que se suponen conocidos) de la información auxiliar. pik, es el vector de probabilidades de inclusión en los elementos incluidos en la muestra. b0, que por defecto toma el valor FALSE indicando que el modelo fue propuesto sin intercepto. De otra forma, si el modelo propuesto contiene intercepto, b0 debe tomar el valor TRUE. El último argumento de la función es ck que hace alusión a la estructura de varianza del modelo. ck toma el valor 1 por defecto. Si la estructura de varianza es como en el modelo de razón, entonces ck deberá ser el mismo vector que se introdujo en el argumento x.

De esta manera, se utiliza la función Wk del paquete TeachingSampling para encontrar los pesos de calibración. Nótese que como el modelo fue propuesto con intercepto, eso quiere decir que la primera columna de la matriz de diseño es de sólo unos; por lo tanto, el argumento tx debe ser un vector conteniendo el total poblacional y el total de la variable de información auxiliar, así tx=c(5,236). Como la estructura de varianza es constante, ck toma el valor uno.

```
> w <- Wk(x.s,tx=c(5,236),pik.s,ck=1,b0=TRUE)
> w
     [,1]     [,2]   [,3]    [,4]
[1,] 0.48    -0.84   5.32    0.04
```

De esta manera se obtienen los pesos calibrado cuya agradable propiedad es que reproducen el total poblacional exacto de la información auxiliar.

```
> sum(x.s*w)                > sum(y.s*w)
[1] 236                     [1] 381.72
```

Sin embargo, si el modelo ξ hubiese sido formulado de manera distinta, como por ejemplo:

$$Y_k = \beta_1 X_k + \varepsilon_k$$

Donde cada uno de los ε_k $k \in U$ son variables aleatorias independientes e idénticamente distribuidas con media cero y estructura de varianza constante. Entonces, los argumentos en la función Wk del paquete TeachingSampling deben cambiar, de tal forma que

```
> w <- Wk(x.s,tx=236,pik.s,ck=1,b0=FALSE)
> w
         [,1]        [,2]        [,3]        [,4]
[1,] 1.581235    1.552844    1.685338    1.571771
```

Nótese que aunque el modelo cambie, la propiedad de calibración se mantiene ante distintas configuraciones en los pesos.

```
> sum(x.s*w)                > sum(y.s*w)
[1] 236                     [1] 380.5163
```

Para este modelo de super-población, haga un ejercicio léxico-gráfico de todas las posibles muestras aleatorias simples de tamaño $n = 4$, donde calcule los pesos de calibración y verifique la propiedad de representatividad sobre el vector de información auxiliar.

9.2.2 Otras propiedades del estimador general de regresión

Por otro lado, acudiendo a la definición del estimador general de regresión, éste toma la siguiente forma

$$\begin{aligned}\hat{t}_{y,greg} &= \sum_U \mathbf{x}'_k\hat{\mathbf{B}} + \sum_s \frac{y_k - \mathbf{x}'_k\hat{\mathbf{B}}}{\pi_k}\\ &= \sum_U \hat{y}_k + \sum_s \frac{e_k}{\pi_k}\end{aligned}$$

En algunas ocasiones, el modelo ξ que establece la relación entre la característica de interés y la información auxiliar es tal que

$$\sum_s \frac{e_k}{\pi_k} = 0.$$

Si la anterior ecuación se satisface, entonces el estimador general de regresión tomaría una forma mucho más sencilla dada por

$$\hat{t}_{y,greg} = \sum_U \hat{y}_k \tag{9.2.25}$$

$$= \sum_U \mathbf{x}'_k\hat{\mathbf{B}} \tag{9.2.26}$$

$$= \mathbf{t}'_{\mathbf{x}}\hat{\mathbf{B}} \tag{9.2.27}$$

Por lo que sólo se necesitaría del conocimiento del vector de totales poblacionales de las variables de información auxiliar $\mathbf{t_x}$, que pueden estar disponibles en alguna entidad administrativa, y de los valores que toman la característica de interés y el vector de información auxiliar, y_k y $\mathbf{x}_k$ respectivamente, en la muestra realizada.

Resultado 9.2.2. *Una condición suficiente para que*

$$\sum_s \frac{e_k}{\pi_k} = 0.$$

es que exista un vector $\mathbf{v}$ *tal que*

$$\mathbf{v}'\mathbf{x}_k = c_k. \tag{9.2.28}$$

Prueba. Si la ecuación (9.2.21) se satisface, entonces

$$\begin{aligned}
\sum_S \frac{e_k}{\pi_k} &= \sum_S \frac{1}{\pi_k}\left(y_k - \mathbf{x}_k'\hat{\mathbf{B}}\right)\\
&= \sum_S \frac{1}{\pi_k}\left(y_k - \frac{\mathbf{v}'\mathbf{x}_k}{c_k}\mathbf{x}_k'\hat{\mathbf{B}}\right)\\
&= \hat{t}_{y,\pi} - \mathbf{v}'\left(\sum_S \frac{\mathbf{x}_k\mathbf{x}_k'}{\pi_k c_k}\right)\hat{\mathbf{T}}^{-1}\widehat{\mathbf{t}}\\
&= \hat{t}_{y,\pi} - \sum_S \frac{\mathbf{v}'\mathbf{x}_k y_k}{\pi_k c_k}\\
&= \hat{t}_{y,\pi} - \hat{t}_{y,\pi} = 0
\end{aligned}$$

■

Särndal, Swensson & Wretman (1992) afirman que algunos ejemplos de estructuras de varianza que satisfacen la ecuación (9.2.21) son:

- Modelo de regresión lineal con intercepto $x_{1k} = 1\ \forall k \in U$ y estructura de varianza constante $c_k = 1$.
- Modelo de regresión lineal con estructura de varianza proporcional a alguna variable del vector de información auxiliar. Es decir,

  $$\sigma^2 c_k \propto x_{jk}$$

  Para algún $j = 1, \ldots, p$ y para todo $k \in U$
- Modelo de regresión lineal con estructura de varianza proporcional a una combinación lineal de las variables de información auxiliar. Es decir,

  $$\sigma^2 c_k \propto \sum_{j=1}^{p} a_j x_{jk}$$

  Para todo $k \in U$ y algunas constantes $a_1, \ldots, a_p$

Acerca de la filosofía que cubre el modelo ξ en el estimador de regresión, Särndal, Swensson & Wretman (1992) afirman que el papel que juega este modelo se limita a la descripción, mas no explicación, de la nube de puntos en la población finita. Argumentan que se espera que el modelo propuesto ajuste razonablemente bien y que haga pensar que pudo haber generado el comportamiento particular de la característica de interés. Nótese que el supuesto es flexible y no exige la certeza de que el modelo en verdad haya generado los valores de y. Por tanto, aunque el modelo induce aleatoriedad per se, las conclusiones de las estimaciones son independientes del mismo. Aún más, el modelo ξ es un vehículo para encontrar una expresión matemática que permita estimar los coeficientes de regresión y la eficiencia de $\hat{t}_{y,greg}$ comparada con la del estimador de Horvitz-Thompson dependerá de la bondad del ajuste inducida por el modelo supuesto. Sin embargo, no depende de

ninguna manera, de si el modelo es cierto o no. Por tanto todo tipo de inferencias acerca del estimador están basados en el diseño de muestreo y no en el modelo supuesto.

Bajo la anterior argumentación, es necesario calcular y estimar la varianza del estimador general de regresión desde un punto de vista basado en el diseño de muestreo. Así que, siguiendo los lineamentos de la sección 8.1.1. en cuanto a la técnica de linealización de Taylor, se tiene el siguiente resultado.

Resultado 9.2.3. *El estimador general de regresión es aproximadamente insesgado para el total poblacional de la característica de interés t_y. Además la aproximación de la varianza y la varianza estimada del estimador general de regresión están dadas por*

$$AVar(\hat{t}_{y,greg}) = \sum\sum_U \Delta_{kl} \frac{E_k}{\pi_k}\frac{E_l}{\pi_l}. \tag{9.2.29}$$

$$\widehat{Var}(\hat{t}_{y,greg}) = \sum\sum_S \frac{\Delta_{kl}}{\pi_{kl}} \frac{e_k}{\pi_k}\frac{e_l}{\pi_l} \tag{9.2.30}$$

respectivamente. Donde $E_k = y_k - \mathbf{x}_k'\mathbf{B}$ son los errores en la población finita y $e_k = y_k - \mathbf{x}_k'\hat{\mathbf{B}}$ son los errores en la muestra seleccionada.

Prueba. Siguiendo los pasos de la linealización de Taylor, debemos expresar el estimador como una función de totales.

$$\begin{aligned} \hat{t}_{y,greg} &= \hat{t}_{y\pi} + (\mathbf{t_x} - \hat{\mathbf{t}}_{\mathbf{x}\pi})'\hat{\mathbf{B}} & (9.2.31)\\ &= f(\hat{t}_{y\pi}, \hat{\mathbf{t}}_{\mathbf{x}\pi}, \hat{\mathbf{T}}, \hat{\mathbf{t}}) & (9.2.32)\end{aligned}$$

Nótese que

$$\left.\frac{\partial f}{\partial \hat{\mathbf{T}}}\right|_{\hat{t}_{y\pi}=t_y, \hat{\mathbf{t}}_{\mathbf{x}\pi}\mathbf{t_x}, \hat{\mathbf{T}}=\mathbf{T}, \hat{\mathbf{t}}=\mathbf{t}} = (\mathbf{t_x} - \hat{\mathbf{t}}_{\mathbf{x}\pi})' \left.\frac{\partial \hat{\mathbf{B}}}{\partial}\right|_{\hat{t}_{y\pi}=t_y, \hat{\mathbf{t}}_{\mathbf{x}\pi}\mathbf{t_x}, \hat{\mathbf{T}}=\mathbf{T}, \hat{\mathbf{t}}=\mathbf{t}} = \mathbf{0}$$

y análogamente, se tiene que

$$\left.\frac{\partial f}{\partial \hat{\mathbf{t}}}\right|_{\hat{t}_{y\pi}=t_y, \hat{\mathbf{t}}_{\mathbf{x}\pi}=\mathbf{t_x}, \hat{\mathbf{T}}=\mathbf{T}, \hat{\mathbf{t}}=\mathbf{t}} = (\mathbf{t_x} - \hat{t}_{\mathbf{x}\pi})' \left.\frac{\partial \hat{\mathbf{B}}}{\partial}\right|_{\hat{\mathbf{t}}_{y\pi}=t_y, \hat{\mathbf{t}}_{\mathbf{x}\pi}=\mathbf{t_x}, \hat{\mathbf{T}}=\mathbf{T}, \hat{\mathbf{t}}=\mathbf{t}} = \mathbf{0}.$$

Por consiguiente, sólo se calcula las derivadas de f con respecto a $\hat{t}_{y\pi}$ y $\hat{\mathbf{t}}_{\mathbf{x}\pi}$, y se tiene que

$$a_1 = \frac{\partial f(\hat{t}_{y\pi}, \hat{\mathbf{t}}_{\mathbf{x}\pi})}{\partial \hat{t}_{y\pi}}\Bigg|_{\hat{t}_{y\pi}=t_y,\hat{\mathbf{t}}_{\mathbf{x}\pi}=\mathbf{t}_\mathbf{x},\hat{\mathbf{T}}=\mathbf{T},\hat{\mathbf{t}}=\mathbf{t}}$$
$$= 1$$
$$a_2 = \frac{\partial f(\hat{t}_{y\pi}, \hat{\mathbf{t}}_{\mathbf{x}\pi})}{\partial \hat{\mathbf{t}}_{\mathbf{x}\pi}}\Bigg|_{\hat{t}_{y\pi}=t_y,\hat{\mathbf{t}}_{\mathbf{x}\pi}=\mathbf{t}_\mathbf{x},\hat{\mathbf{T}}=\mathbf{T},\hat{\mathbf{t}}=\mathbf{t}}$$
$$= -\hat{\mathbf{B}}'\Big|_{\hat{t}_{y\pi}=t_y,\hat{\mathbf{t}}_{\mathbf{x}\pi}=\mathbf{t}_\mathbf{x},\hat{\mathbf{T}}=\mathbf{T},\hat{\mathbf{t}}=\mathbf{t}}$$
$$= -\mathbf{B}'$$

Por tanto, se tiene que

$$\hat{t}_{y,greg} \cong t_y + (\hat{t}_{y\pi} - t_y) - \mathbf{B}'(\hat{\mathbf{t}}_{\mathbf{x}\pi} - \hat{\mathbf{t}}_\mathbf{x}),$$

y tomando esperanza, se tiene que $E(\hat{t}_{y,greg}) \cong t_y$.
Al definir la nueva variable linealizada dada por la expresión (8.1.14), se tiene que

$$E_k = y_k - \mathbf{x}_k'\mathbf{B} \tag{9.2.33}$$

cuya aproximación con los datos recolectados en la muestra es

$$e_k = y_k - \mathbf{x}_k'\widehat{\mathbf{B}} \tag{9.2.34}$$

Por tanto, la varianza se escribe, recurriendo al resultado 8.1.3, como

$$AVar(\hat{t}_{y,greg}) = Var\left(\sum_S \frac{E_k}{\pi_k}\right) \tag{9.2.35}$$

Utilizando los principios del estimador de Horvitz-Thompson se llega a los resultados de la aproximación de la varianza y de la varianza estimada. ■

Särndal, Swensson & Wretman (1992) proponen un estimador de la varianza que integra los pesos g_{ks}. La motivación de este nuevo estimador de la varianza recae en que una forma de escribir el estimador de regresión general está dada por

$$\hat{t}_{y,greg} = \sum_U y_k{}^{\circ} + \sum_S \frac{g_{ks}E_k}{\pi_k} \tag{9.2.36}$$

Por lo tanto, al calcular su varianza tenemos

$$Var\left(\hat{t}_{y,greg}\right) = Var\left(\sum_U y_k{}^{\circ} + \sum_S \frac{g_{ks}E_k}{\pi_k}\right) \tag{9.2.37}$$
$$= Var\left(\sum_S \frac{g_{ks}E_k}{\pi_k}\right) \tag{9.2.38}$$

Utilizando los principios del estimador de Horvitz-Thompson, un estimador alternativo para la varianza del estimador general de regresión está dada por

$$\widehat{Var}(\hat{t}_{y,greg}) = \sum\sum_S \frac{\Delta_{kl}}{\pi_{kl}} \frac{g_{ks}e_k}{\pi_k} \frac{g_{ls}e_l}{\pi_l} \tag{9.2.39}$$

El lector debe tener muy claro que la propiedad de insesgamiento no aplica a esta clase de estimadores. Sin embargo, cuando el tamaño de muestra y el tamaño poblacional son grandes, entonces el sesgo del estimador general de regresión es despreciable. Se debe tener sumo cuidado en las muestras de tamaño pequeño, máxime cuando se realiza el proceso de estimación por intervalos de confianza. Särndal, Swensson & Wretman (1992) afirman al respecto que, aunque el sesgo afecta la validez de los intervalos de confianza generados con el estimador general de regresión, es válido utilizar el siguiente intervalo de confianza

$$\hat{t}_{y,greg} \pm z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{t}_{y,greg})} \tag{9.2.40}$$

incluso cuando el tamaño de muestra es modesto.

Resultado 9.2.4. *Bajo la familia de diseños de muestreo MAS, el estimador general de regresión es consistente en el sentido Cochran. Es decir, si $s = U$, entonces*

$$\hat{t}_{y,greg} = t_y \tag{9.2.41}$$

Hasta este momento, hemos definido el estimador general de regresión como un intento de conciliar la teoría clásica de modelos con el muestreo de poblaciones finitas. Este estimador ha ganado mucho campo en las últimas décadas y su uso, dadas sus propiedades deseables, es aún mayor a medida que el tiempo pasa. Sin embargo, el estimador general de regresión es el resultado de décadas de desarrollo teórico y construcción de estimadores asistidos por modelos que se constituyen como casos particulares de éste.

En las próximas secciones, estudiaremos cada uno de estos casos particulares más utilizados en la práctica. El lector debe notar que cada uno de los estimadores que siguen en las siguientes secciones, fueron propuestos en los tiempos antiguos sin tener en cuenta un modelo de super-población sino con una motivación puramente empírica. Sin embargo, como se verá en desarrollo de las siguientes secciones, todos estos estimadores están cubiertos bajo los principios del estimador general de regresión y por los coeficientes de regresión que el modelo induzca.

Para terminar la exposición del estimador general de regresión, el lector debe notar que este estimador es completamente inútil en la práctica. En otras palabras, su basta generalidad hace que este estimador sea inutilizable. Como en todo proceso estadístico, el modelo general y sus correspondientes expresiones matemáticas carecen de sentido sin el conocimiento del comportamiento particular de cada característica de interés. Con lo anterior, no es mi intención desactivar al lector. Por el contrario, cuando el estadístico logra entender qué es un modelo de super población, y obtiene un estimador particular conforme al comportamiento de la población de estudio, entonces la ganancia en eficiencia es tremenda.

En las siguientes secciones se darán ejemplos particulares del estimador de regresión cuando el modelo que rige la población finita ya se ha especificado. Nótese

que todos y cada uno de los estimadores que a continuación se presentan son casos particulares del estimador general de regresión. Por supuesto, cada uno de ellos recibe un nombre particular, que en la mayoría de los casos está supeditado al modelo que rige la población particular.

El lector debe retomar en cada una de las siguientes páginas el espíritu del estimador general de regresión como una familia que cobija casos particulares de estimadores. Todos y cada uno de los estimadores que se revisan en este capítulo nacieron bajo especificaciones propias que los caracterizaban de manera singular. Por tanto, el desarrollo histórico de cada uno de ellos no estuvo fundamentado, en principio, como un caso particular de algún otro estimador. El estimador de razón, el estimador de regresión, el estimador de post-estratificación, entre otros, fueron concebidos aparte de la idea de los modelos lineales. Sus creadores no estaban pensando en calcular o estimar un coeficiente de regresión. Por supuesto, con el transcurrir del tiempo y los avances en términos de la teoría estadística de los modelos lineales, se creó una familia que unifica a todos los estimadores de este capítulo en un sólo estimador general.

9.3 Estimador de media común

Recuerde que la construcción de la estrategia de muestreo es la tarea más importante antes de realizar cualquier estudio por muestreo. Sin embargo, se debe reconocer que cada una de las posibles estrategias de muestreo tiene ventajas y desventajas sobre las restantes estrategias. Suponga que el diseño de muestreo que se ha propuesto consiste en un diseño de muestreo Bernoulli. ¿Qué tipo de estimador es el mejor para este diseño de muestreo?. En teoría, existen muchos estimadores insesgados para este diseño particular, por ejemplo el estimador de Horvitz-Thompson. Sin embargo, desde un punto de vista práctico, es posible que la muestra realizada o seleccionada para este diseño de muestreo consista en todas y cada una de las unidades de la población. Bajo el anterior escenario el estimador de Horvitz-Thompson no plantea ningún tipo de ventajas pues la estimación para el total poblacional será una estimación totalmente errónea, igual a t_y/π y estrictamente mayor a t_y.

Como se vio en capítulos anteriores, aunque la probabilidad de que la muestra seleccionada o realizada contenga todas las unidades poblacionales, el estimador alternativo del total poblacional, dado en la expresión (2.2.17), proporciona una mejor opción que el estimador de Horvitz-Thompson. Este estimador alternativo se conoce con el nombre de estimador de media común y está motivado por el **modelo de media común** que supone que la población se comporta de la misma manera de acuerdo a una pendiente común para cada uno de los individuos que conforman. De esta manera $p = 1$, $\mathbf{x}_k = 1$ y $c_k = 1$ para todo $k \in U$. La formulación del modelo de superpoblación está dada por

$$Y_k = \beta + \varepsilon_k \tag{9.3.1}$$

Donde cada uno de los ε_k $k \in U$ son variables aleatorias independientes e idénticamente distribuidas con media cero y varianza σ^2. Como resultado de lo anterior se tiene que

$$
\begin{aligned}
E_\xi(Y_k) &= \beta \\
Var_\xi(Y_k) &= \sigma^2.
\end{aligned}
\tag{9.3.2}
$$

A simple vista el estimador resultante del modelo anterior no es mejor que el estimador de Horvitz-Thompson pues la información auxiliar es siempre constante. Sin embargo, el estimador resultante es muchas veces mejor que el estimador de Horvitz-Thompson como cuando la estrategia de muestreo implica un diseño de muestreo tipo Bernoulli. Es común utilizar el estimador de media común cuando el gráfico de dispersión entre la característica de interés y la característica de información auxiliar define una recta de regresión constante y paralela al eje de las abscisas. Por supuesto, el cociente entre estas dos características también definirá un gráfico de dispersión cuyo comportamiento sea constante con ligeras desviaciones uniformes como se puede observar en la siguiente figura.

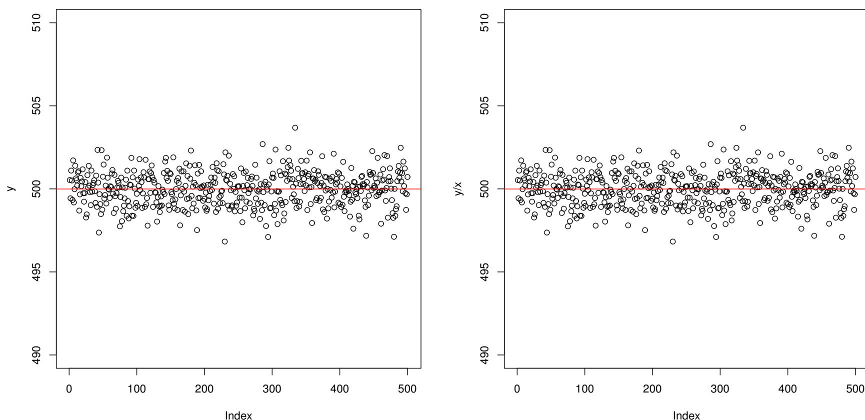


Figura 9.1: *Relación en un modelo de media común.*

Si se tuviese acceso a toda la población finita, el estimador del coeficiente de regresión β estaría dado por la minimización de la siguiente función de dispersión

$$D = \sum_U \frac{(y_k - B)^2}{\sigma^2}. \tag{9.3.3}$$

Utilizando el resultado 8.4.2 y recurriendo a la ecuación (8.4.6), el estimador B en la población finita toma la siguiente forma

$$B = \frac{t_y}{N} = \bar{y}_U \tag{9.3.4}$$

Por supuesto, como en la práctica sólo se tiene acceso a una muestra particular de población finita, B debe ser estimado de tal manera que siguiendo el resultado 8.4.3. llegamos a la siguiente expresión

$$\hat{B} = \frac{\hat{y}_{y,\pi}}{\hat{N}_\pi} = \widetilde{y}_S \tag{9.3.5}$$

Con estas herramientas es posible ahora construir un estimador del total poblacional de la característica de interés el cual está dado por el siguiente resultado.

Resultado 9.3.1. *Bajo el modelo de media común, el estimador del total poblacional está dado por*

$$\hat{t}_{y,mc} = N\frac{\hat{t}_{y,\pi}}{\hat{N}_{\pi}} = N\widetilde{y}_S \tag{9.3.6}$$

cuya varianza aproximada es

$$AVar(\hat{t}_{y,mc}) = \sum\sum_U \Delta_{kl}\frac{E_k}{\pi_k}\frac{E_l}{\pi_l}. \tag{9.3.7}$$

con

$$E_k = y_k - B \tag{9.3.8}$$

$$= y_k - \frac{t_y}{N} = y_k - \bar{y}_U. \tag{9.3.9}$$

El estimador de la varianza es

$$\widehat{Var}(\hat{t}_{y,greg}) = \sum\sum_S \frac{\Delta_{kl}}{\pi_{kl}}\frac{e_k}{\pi_k}\frac{e_l}{\pi_l} \tag{9.3.10}$$

con

$$e_k = y_k - \hat{B} \tag{9.3.11}$$

$$= y_k - \frac{\hat{t}_{y,\pi}}{\hat{N}_{\pi}} = y_k - \widetilde{y}_S. \tag{9.3.12}$$

Prueba. Antes de empezar la demostración, el lector debe tener en cuenta que estimador es un caso particular del estimador general de regresión. Por lo tanto, como $\mathbf{x}_k = 1$ para todo $k \in U$, adecuando la expresión (9.2.11) se tiene que

$$\hat{t}_{y,mc} = \hat{t}_{y,\pi} + \hat{B}(t_x - \hat{t}_{x,\pi}) \tag{9.3.13}$$

$$= \hat{t}_{y,\pi} + \frac{\hat{t}_{y,\pi}}{\hat{N}_{\pi}}(N - \hat{N}_{\pi}) \tag{9.3.14}$$

$$= N\frac{\hat{t}_{y,\pi}}{\hat{N}_{\pi}} = N\widetilde{y}_S \tag{9.3.15}$$

El cálculo de la varianza aproximada y la estimación de la varianza del estimador de razón son inmediatos al utilizar el resultado 9.2.3. ■

El espíritu y la ventaja de este estimador está en la corrección que hace al estimador de Horvitz-Thompson mediante el cociente $\frac{N}{\hat{N}_{\pi}}$. De esta manera, cuando el estimador de Horvitz-Thompson está subestimando o sobreestimando el total poblacional, entonces este cociente corrige inmediatamente esta sub o sobre estimación.

A continuación se presentan otras características importantes del estimador de media común para el total poblacional. En primer lugar, nótese que fácilmente se puede demostrar que

$$\sum_s \frac{e_k}{\pi_k} = 0$$

Lo anterior se tiene puesto que, recurriendo al resultado 9.2.2, $\mathbf{x}_k = c_k = 1$ y por lo tanto $\mathbf{v}' = 1$. Como consecuencia de lo anterior, es posible escribir al estimador de media común en una forma simplificada

$$\hat{t}_{y,mc} = \sum_U \hat{y}_k = \sum_U \hat{B} \tag{9.3.16}$$

$$= \sum_U \widetilde{y}_S = N\widetilde{y}_S \tag{9.3.17}$$

Además recurriendo a las expresiones (9.2.16) y (9.2.17) se tiene que

$$g_{ks} = 1 + \left(t_x - \hat{t}_{x,\pi}\right)\left(\hat{t}_{x,\pi}\right)^{-1} \tag{9.3.18}$$

$$= 1 + \left(\frac{N - \hat{N}_\pi}{\hat{N}_\pi}\right) = \frac{N}{\hat{N}_\pi} \tag{9.3.19}$$

9.3.1 Algunos diseños de muestreo

Diseño de muestreo Bernoulli

Bajo el diseño de muestreo Bernoulli, el estimador de media común toma una forma idéntica al estimador alternativo propuesto en la expresión (3.1.14) de la sección 3.1. En esos apartados, no se dieron las expresiones para la varianza y la varianza estimada puesto que se requería de herramientas de las que no se disponían. Sin embargo, el siguiente resultado da cuenta de las expresiones exactas para este estimador alternativo.

Resultado 9.3.2. *Si el diseño de muestreo es Bernoulli, el estimador de media común, su varianza aproximada y el estimador de la varianza están dados por*

$$\hat{t}_{y,mc} = N\widetilde{y}_S = N\frac{\sum_S y_k}{n(S)} = N\bar{y}_S. \tag{9.3.20}$$

$$AV_{BER}\hat{t}_{y,mc} = N\left(\frac{1}{\pi} - 1\right)S^2_{y_U} \tag{9.3.21}$$

$$\hat{Var}_{BER}\hat{t}_{y,mc} = (n(S) - 1)\frac{1}{\pi}\left(\frac{1}{\pi} - 1\right)S^2_{y_S} \tag{9.3.22}$$

respectivamente. Con $S^2_{y_U}$ la varianza poblacional de la característica de interés y $S^2_{y_S}$ la varianza muestral de la característica de interés.

Prueba. El resultado se sigue inmediatamente al evaluar la expresión (3.1.12) en cada una de las ecuaciones del resultado. ■

Diseño de muestreo aleatorio simple

Resultado 9.3.3. *Si el diseño de muestreo es aleatorio simple, el estimador de media común toma la misma forma que el estimador de Horvitz-Thompson. Por supuesto, la varianza aproximada y el estimador de la varianza son los mismos que los del estimador de Horvitz-Thompson. En general, se tiene que*

$$\hat{t}_{y,mc} = N\widetilde{y}_S = \frac{N}{n}\sum_S y_k \tag{9.3.23}$$

$$Var_{MAS}(\hat{t}_{y,mc}) = \frac{N^2}{n}\left(1-\frac{n}{N}\right)S^2_{E_U} \tag{9.3.24}$$

$$\widehat{Var}_{MAS}(\hat{t}_{y,mc}) = \frac{N^2}{n}\left(1-\frac{n}{N}\right)S^2_{e_S} \tag{9.3.25}$$

respectivamente. Con $S^2_{E_U}$ la varianza poblacional de los errores $E_k = y_k - \bar{y}_U$ y $S^2_{e_S}$ la varianza muestral de los errores $e_k = y_k - \bar{y}_S$.

Prueba. El resultado se sigue inmediatamente al aplicar los principios del estimador de Horvitz-Thompson a las expresiones (9.3.7) y (9.3.10) bajo el diseño de muestreo aleatorio simple. Nótese que bajo el diseño de muestreo aleatorio simple, $\bar{E} = 0$ y $\bar{e} = 0$, por lo tanto $S^2_{E_U} = S^2_{y_U}$ y $S^2_{e_S} = S^2_{y_S}$. ■

9.3.2 Marco y Lucy

Retomando la población de empresas pertenecientes al sector industrial, suponga que se desea estimar el total de las características de interés mediante un estimador de regresión que obedezca al modelo dado por la expresión (9.3.2), en donde las características de interés están relacionadas con una variable que es constante y que supone el mismo comportamiento estructural a lo largo de toda la población. Suponga que se selecciona una muestra aleatoria simple de tamaño $n = 400$

```
> data(Lucy)
> attach(Lucy)
> N <- dim(Lucy)[1]
> n <- 400
> sam <- S.SI(N,n)
> muestra <- Lucy[sam,]
> attach(muestra)
```

Para computar el estimador del total de las características de interés se define la matriz de información auxiliar, que en este caso particular corresponde a un vector de unos y se utiliza la función `GREG.SI` del paquete `TeachingSampling` que cuenta con siete argumentos: `N`, el tamaño poblacional, `n`, el tamaño de la muestra, `y`, correspondiente al vector o matriz de datos que contienen las observaciones de los individuos incluidos en la muestra, `x`, concerniente al vector o matriz de información auxiliar en la muestra, `tx`, el total poblacional de las variables de

información auxiliar, b, el estimador de coeficientes de regresión y, por último, b0, que indica si el modelo está definido con o sin intercepto.

Por consiguiente, definiendo correctamente los parámetros según el modelo dado por (9.3.2), tenemos el siguiente código computacional para el cálculo del estimador del total poblacional.

```
> estima <- data.frame(Income, Employees, Taxes)
> pik <- rep(n/N,n)
> x <- rep(1,n)
> tx <- c(N)
> b <- E.Beta(estima,x,pik,ck=1,b0=FALSE)
> GREG.SI(N,n,estima,x,tx, b, b0=FALSE)
                               Income    Employees        Taxes
Total estimado           1.013059e+06 1.502891e+05 2.759294e+04
Varianza                 8.305089e+08 1.268017e+07 3.360041e+06
coeficiente de variación 2.844707e+00 2.369382e+00 6.643155e+00
```

Por otra parte, el siguiente código muestra que la propiedad de calibración se mantiene para este modelo. Para esto, se planea la utilización de la función Wk del paquete TeachingSampling que cuenta con cinco argumentos: x, concerniente al vector o matriz de información auxiliar en la muestra, tx, el total poblacional de las variables de información auxiliar, pik, el vector de probabilidades de inclusión para los elementos de la muestra, *ck*, que hace alusión a la estructura de varianza del modelo y, por último b0 que se refiere al intercepto del modelo. Nótese que la multiplicación de los valores observados de cada una de las características de interés con los pesos de calibración da como resultado la misma estimación que al utilizar la función GREG.SI.

```
> w <- Wk(x,tx,pik,ck=1,b0=FALSE)
> sum(x*w)
[1] 2396
>  colSums(w*estima)
      Income    Employees          Taxes
1.013059e+06 1.502891e+05  2.759294e+04
```

A continuación se muestra la tabla de resultados que indica que asumiendo este modelo se gana mucha más precisión que en el caso en donde no se asume ningún modelo y se realizan las estimaciones directamente como en el caso del diseño de muestreo aleatorio simple. Nótese que para este caso particular, no se tuvo a la mano ninguna característica de información auxiliar sino que el modelo estuvo inducido por un simple vector de constantes que mejoro la estimación y la precisión.

9.4 Estimador de razón

El estimador de razón fue el primer intento de conciliar la información auxiliar en la etapa de estimación bajo un marco de inferencia basado en el diseño para mejorar la eficiencia del estimador del total poblacional y es atribuido a Hájek (1971) aunque siglos atrás Laplace (1814) lo utilizó para estimar el total de habitantes en Francia.

Tabla 9.1: *Modelo de media común: estimación de los totales de las características de interés.*

Variable	Total poblacional	Total estimado	cve %	Desv. %
Ingreso	1035217	1013059	2.84	-2.14
Empleados	151950	150289	2.36	-1.09
Impuestos	28654	27592	6.64	-3.70

Suponga que existe una variable de información auxiliar continua x, entonces para poder utilizar un estimador de razón es necesario contar con los valores de la información auxiliar en la muestra x_k $k \in S$ y con el total poblacional de la misma, t_x. Esta última información puede ser obtenida de fuentes administrativas o incluso estadísticas oficiales manejadas por los institutos de estadística.

El modelo existente detrás de la construcción de este estimador está dado por el **modelo de razón** que supone la existencia de una sola variable de información auxiliar continua relacionada con la característica de interés, cuya estructura de varianza es inversamente proporcional al comportamiento estructural de la información auxiliar. Así que $p = 1$, $\mathbf{x}_k = x_k$ y $c_k = x_k$ para todo $k \in U$. La formulación del modelo está dada por

$$Y_k = \beta X_k + \varepsilon_k \tag{9.4.1}$$

Donde cada uno de los ε_k $k \in U$ son variables aleatorias independientes e idénticamente distribuidas con media cero y varianza $x_k\sigma^2$. En donde,

$$\begin{aligned} E_\xi(Y_k) &= \beta x_k \\ Var_\xi(Y_k) &= x_k\sigma^2. \end{aligned} \tag{9.4.2}$$

Lohr (2000) plantea que la utilización del estimador de razón es más adecuada cuando la recta que resume la relación entre la característica de interés y la información auxiliar pasa por el origen y si la varianza de y_k es proporcional a x_k, es decir a medida que la característica de interés toma valores más grandes, la dispersión se hace mucho mayor, justo como se observa en la siguiente figura en la que se presenta el gráfico de dispersión entre las variables del modelo y también la dispersión del cociente de las dos variables, en donde la relación es muy constante.

De esta forma, B corresponde a la pendiente de la recta de regresión ponderada proporcionalmente a $1/x_k$,hallada mediante el método de mínimos cuadrados, que minimiza la siguiente función:

$$D = \sum_U \left(\frac{y_k - Bx_k}{x_k\sigma^2} \right)^2. \tag{9.4.3}$$

Mediante la expresión (8.4.6), no es difícil mostrar que bajo el modelo de razón, la estimación de β en la población finita es

$$B = \frac{t_y}{t_x} \tag{9.4.4}$$

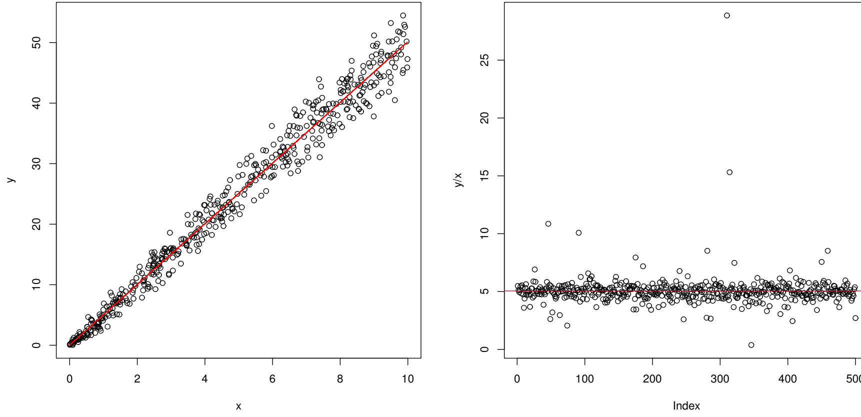


Figura 9.2: *Relación en un modelo de razón.*

Sin embargo, en las aplicaciones prácticas, sólo tenemos acceso a los valores de los elementos seleccionados en la muestra; por tanto, siguiendo el resultado 8.4.3, la estimación de B es

$$\hat{B} = \frac{\hat{t}_{y,\pi}}{\hat{t}_{x,\pi}} \tag{9.4.5}$$

Resultado 9.4.1. *Bajo el modelo de razón, el estimador del total poblacional está dado por*

$$\hat{t}_{y,r} = t_x \frac{\hat{t}_{y,\pi}}{\hat{t}_{x,\pi}} \tag{9.4.6}$$

cuya varianza aproximada es

$$AVar(\hat{t}_{y,r}) = \sum\sum_U \Delta_{kl} \frac{E_k}{\pi_k} \frac{E_l}{\pi_l}. \tag{9.4.7}$$

con

$$E_k = y_k - Bx_k \tag{9.4.8}$$
$$= y_k - \frac{t_y}{t_x} x_k. \tag{9.4.9}$$

El estimador de la varianza es

$$\widehat{Var}(\hat{t}_{y,greg}) = \sum\sum_S \frac{\Delta_{kl}}{\pi_{kl}} \frac{e_k}{\pi_k} \frac{e_l}{\pi_l} \tag{9.4.10}$$

con

$$
\begin{aligned}
e_k &= y_k - \hat{B}x_k & (9.4.11)\\
&= y_k - \frac{\hat{t}_{y,\pi}}{\hat{t}_{x,\pi}} x_k. & (9.4.12)
\end{aligned}
$$

Prueba. En este punto, el lector debe tener claridad de que el estimador de razón es un caso particular del estimador general de regresión. En ese orden de ideas, partiendo de la expresión (9.2.11) se tiene que

$$
\begin{aligned}
\hat{t}_{y,r} &= \hat{t}_{y,\pi} + \hat{B}(t_x - \hat{t}_{x,\pi}) & (9.4.13)\\
&= \hat{t}_{y,\pi} + \frac{\hat{t}_{y,\pi}}{\hat{t}_{x,\pi}}(t_x - \hat{t}_{x,\pi}) & (9.4.14)\\
&= t_x \frac{\hat{t}_{y,\pi}}{\hat{t}_{x,\pi}} & (9.4.15)
\end{aligned}
$$

El cálculo de la varianza aproximada y la estimación de la varianza del estimador de razón son inmediatos al utilizar el resultado 9.2.3. ■

Nótese que el estimador de razón hace uso de la información auxiliar de manera correctiva. Es decir, la expresión $\frac{t_x}{\hat{t}_{x,\pi}}$ hace un ajuste en la estimación de Horvitz-Thompson del total de la característica auxiliar y, por lo tanto, la corrige. Un aspecto importante del estimador de razón es que verifica el resultado 9.2.2 fácilmente haciendo $\mathbf{v}' = \sigma^2$, por lo tanto

$$
\sum_s \frac{e_k}{\pi_k} = 0
$$

y es posible escribir el estimador de razón en su forma simplificada.

$$
\begin{aligned}
\hat{t}_{y,r} &= \sum_U \hat{y}_k & (9.4.16)\\
&= \sum_U \hat{B}x_k & (9.4.17)
\end{aligned}
$$

Además, de la expresión (9.2.17) se tiene que

$$
\begin{aligned}
g_{ks} &= 1 + \left(t_x - \hat{t}_{x,\pi}\right)\left(\hat{t}_{x,\pi}\right)^{-1} & (9.4.18)\\
&= \frac{t_x}{\hat{t}_{x,\pi}} & (9.4.19)
\end{aligned}
$$

Aunque $\widehat{B}$ corresponde a un coeficiente de regresión inducido por el modelo ξ, la verdadera motivación detrás del estimador de razón es que si existe la misma

relación entre la característica de interés y la información auxiliar para todos los elementos de la población entonces se habla de la razón poblacional estaría dada por

$$B = \frac{t_y}{t_x}$$

Despejando de la anterior expresión a t_y se tiene que

$$t_y = t_x B$$

Como t_x es conocido, un estimador natural e intuitivo de t_y está dado por

$$\hat{t}_y = t_x \frac{\hat{t}_{y,\pi}}{\hat{t}_{x,\pi}}$$

Aparte de estimar la razón poblacional y utilizar la información auxiliar continua para corregir el estimador de Horvitz-Thompson, es posible utilizar el estimador de razón en los siguientes casos (Lohr 2000).

- Cuando se desconoce N y es necesario estimarlo. Debido a que $N = t_x/\bar{x}$, entonces $\hat{N} = t_x/\hat{\bar{x}}_\pi$.
- Aumentar la precisión de los estimadores de totales y medias poblacionales.
- Ajustar las estimaciones de la muestra para que reflejen los totales demográficos y cuando se presenta ausencia de respuesta.

9.4.1 Algunos diseños de muestreo

Muestreo Bernoulli

Resultado 9.4.2. *Si el diseño de muestreo es Bernoulli, el estimador de razón toma, su varianza aproximada y el estimador de la varianza están dados por:*

$$\hat{t}_{y,r} = t_x \frac{\sum_S y_k}{\sum_S x_k} \tag{9.4.20}$$

$$AVar_{BER}(\hat{t}_{y,r}) = \left(\frac{1}{\pi} - 1\right) \sum_U E_k \tag{9.4.21}$$

$$\widehat{Var}_{BER}(\hat{t}_{y,r}) = \frac{1}{\pi} \left(\frac{1}{\pi} - 1\right) \sum_U E_k \tag{9.4.22}$$

respectivamente.

Muestreo aleatorio simple

Resultado 9.4.3. *Si el diseño de muestreo es aleatorio simple, el estimador de razón toma, su varianza aproximada y el estimador de la varianza están dados por*

$$\hat{t}_{y,r} = t_x \frac{\bar{y}_S}{\bar{x}_S} \tag{9.4.23}$$

$$AVar_{MAS}(\hat{t}_{y,r}) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right) S^2_{E_U} \tag{9.4.24}$$

$$\widehat{Var}_{MAS}(\hat{t}_{y,r}) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right) S^2_{e_S} \tag{9.4.25}$$

respectivamente, con $S^2_{E_U}$ *la varianza poblacional de la variable linealizada* E_k *y* $S^2_{e_s}$ *la varianza muestral de los valores* e_k.

9.4.2 Marco y Lucy

Volviendo una vez más con la población de empresas pertenecientes al sector industrial, suponga que se desea estimar el total de las características de interés Ingreso (real en el último año fiscal) y que se utiliza un estimador de razón que obedezca al modelo dado por la expresión (9.4.2), en donde esta característica de interés está relacionadas con la característica de información auxiliar Número de Empleados. Además se supone que la dispersión es constante a lo largo de toda la población. Suponga que se selecciona una muestra aleatoria simple de tamaño $n = 400$

```
> data(Lucy)
> attach(Lucy)
> N <- dim(Lucy)[1]
> n <- 400
> sam <- S.SI(N,n)
> muestra <- Lucy[sam,]
> attach(muestra)
```

Para computar el estimador del total de la característica de interés se define la matriz de información auxiliar, que en este caso particular corresponde a una vector de valores que corresponden al número de empleados en cada una de las empresas seleccionadas en la muestra y se utiliza la función `GREG.SI` del paquete `TeachingSampling` definiendo correctamente los parámetros según el modelo dado por (9.4.2).

```
> estima <- data.frame(Income)
> pik <- rep(n/N,n)
> x <- Employees
> tx <- c(151950)
> b <- E.Beta(estima,x,pik,ck=x,b0=FALSE)
> GREG.SI(N,n,estima,x,tx, b, b0=FALSE)
                                   Income
```

```
Total estimado            1.059225e+06
Varianza                  5.940520e+08
coeficiente de variación 2.301040e+00
```

El siguiente código muestra que la propiedad de calibración se mantiene para este modelo. Por supuesto que la multiplicación de los valores observados de la característica de interés Ingreso con los pesos de calibración da como resultado la misma estimación que al utilizar la función `GREG.SI`.

```
> w <- Wk(x,tx,pik,ck=x,b0=FALSE)
> sum(x*w)
[1] 151950
> sum(Income*w)
[1] 1066829
```

A continuación se muestra la tabla de resultados que indica que asumiendo este modelo se gana más precisión que en el caso en donde se asume un modelo simple.

Tabla 9.2: *modelo de razón: estimación de los totales de la característica de interés.*

Variable	Total poblacional	Total estimado	cve %	Desv. %
Ingreso	1035217	1059225	2.30	2.31

9.4.3 Muestreo estratificado

Hace una buena cantidad de años la teoría de muestreo era enseñada de una manera totalmente distinta a la de hoy en día, en donde se tienen marcos de referencia para la creación de cualquier estrategia de muestreo. En ese entonces, el contenido temático de un curso de muestreo se enfocaba en repasar cada uno de los diseños de muestreo más utilizados y para cada uno de ellos se estudiaban unos cuantos estimadores que, por la naturaleza de su concepción, no tenían mayor vínculo el uno con el otro. Sin embargo, el espíritu no ha cambiado; puesto que, al igual que ahora, el objetivo era mejorar la eficiencia de la estrategia de muestreo. En particular, cuando le llegaba el turno al diseño de muestreo estratificado, empezaba un desfile de estimadores que tomaban ciertas expresiones particulares para este diseño.

Cuando se trataba de mejorar la eficiencia mediante la incorporación de la información auxiliar en la etapa de estimación, el camino se bifurcaba en dos:

- En primer lugar, es bien sabido que una de las causas de la estratificación es el comportamiento disímil de la característica de interés en cada uno de los estratos propuestos. Siendo así, la incorporación de información auxiliar amerita la incorporación de tantos modelos como estratos considerados. El anterior argumento es lógico puesto que la relación entre la característica de interés y la información auxiliar no tiene por qué ser la misma en cada estrato. Por ejemplo, Bautista (1998) considera que la relación entre producción industrial y personal ocupado es diferente entre empresas de diferente rama de actividad, pero muy parecida entre empresas de la misma rama. Otro

ejemplo que podría citarse es la relación entre dinero invertido en recreación entre la población de distintos niveles socioeconómicos.

- Por otra parte, cuando la estratificación se realiza por motivos, demográficos o logísticos, es viable pensar que existe una característica de información auxiliar que esté relacionada de igual manera, independientemente del estrato, con la característica de interés. Una vez más, Bautista (1998) da el ejemplo de la estimación del total de pacientes admitidos a un hospital. Nótese que es pertinente pensar en que una característica de información auxiliar sea el número de camas en cada uno de los hospitales. De esta forma, la relación es invariante al estrato.

En cada uno de los casos anteriores existen estimadores particulares cuya pertinencia debe ser analizada desde un punto de vista muy crítico y siempre teniendo en cuenta el vínculo con el capítulo de estimadores de coeficientes de regresión, específicamente si existe la necesidad o no de formular un **modelo de grupo**.

Por supuesto, el estimador de razón no es el único estimador con el cual se debe hacer el anterior razonamiento. El estimador general de regresión admite este razonamiento cada vez que la estrategia de muestreo involucre un diseño estratificado. De tal forma que, cuando se considere el primer caso, es decir tantos modelos como estratos, se define el **estimador de regresión separada** y cuando se considere el segundo caso, es decir un sólo modelo para toda la población estratificada, se define el **estimador de regresión combinada**.

El apellido de estos estimadores generales cambia de acuerdo al estimador que se considere. Por ejemplo, si se considera el estimador de razón, se habla entonces del **estimador de razón separada** y del **estimador de razón combinada**.

Muestreo aleatorio estratificado: estimador de razón combinada

Un importante caso particular del estimador de razón, que se rige bajo el modelo dado por la expresión (9.4.2) es el estimador de razón combinada cuyo énfasis es que, a pesar de que la población esté estratificada, sólo involucra un modelo.

Resultado 9.4.4. *Bajo diseño de muestreo aleatorio estratificado, el estimador de razón combinada del total poblacional t_y, su varianza y su varianza estimada están dados por*

$$\hat{t}_{y,rc} = \left(\sum_U x_k\right) \frac{\sum_{h=1}^H N_h \bar{y}_{s_h}}{\sum_{h=1}^H N_h \bar{x}_{s_h}} \tag{9.4.26}$$

$$Var_{MAE}(\hat{t}_{y,rc}) = \sum_{h=1}^{H} \frac{N_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right) S^2_{E_{U_h}} \tag{9.4.27}$$

$$\widehat{Var}_{MAE}(\hat{t}_{y,rc}) = \sum_{h=1}^{H} \frac{N_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right) S^2_{e_{S_h}} \tag{9.4.28}$$

respectivamente. Donde $S^2_{E_{U_h}}$ es la varianza poblacional de la variable linealizada E_k en el estrato h-ésimo y $S^2_{e_{S_h}}$ es la varianza muestral de los valores e_k en el

estrato h-ésimo. Nótese que

$$E_k = y_k - \left(\frac{\sum_{h=1}^{H} N_h \bar{y}_{U_h}}{\sum_{h=1}^{H} N_h \bar{x}_{U_h}}\right) x_k$$

$$e_k = y_k - \left(\frac{\sum_{h=1}^{H} N_h \bar{y}_{S_h}}{\sum_{h=1}^{H} N_h \bar{x}_{S_h}}\right) x_k$$

Prueba. La demostración es inmediata al utilizar los resultados 5.1.3. y 5.1.4 junto con las expresiones (9.4.6), (9.4.7) y (9.4.10). ■

Muestreo aleatorio simple: estimador de razón separada

El estimador de razón separada es un caso especial del estimador de razón pero no comparte el mismo modelo, aunque sí asume que la población se encuentra estratificada y en cada estrato se supone un modelo distinto. Se supone la existencia de una sola variable de información auxiliar continua relacionada con la característica de interés, cuya estructura de varianza es inversamente proporcional al comportamiento estructural de la información auxiliar. Por otra parte se supone también que tanto la característica de interés, como la información auxiliar tienen un comportamiento distinto, y por tanto una relación distinta, en cada uno de los estratos. De esta forma, $p = 1$, $\mathbf{x}_k = x_k$ y $c_k = x_k$ para todo $k \in U_h$ con $h = 1, \ldots, H$. La formulación del modelo, que comúnmente se conoce como **modelo de razón separada**, está dada por

$$Y_k = \beta_h X_k + \varepsilon_k \qquad \text{para todo } h = 1, \ldots, H \tag{9.4.29}$$

Donde cada uno de los ε_k $k \in U_h$ son variables aleatorias independientes e idénticamente distribuidas con media cero y varianza $x_k\sigma_h^2$. En donde,

$$\begin{aligned} E_\xi(Y_k) &= \beta_h x_k \\ Var_\xi(Y_k) &= x_k\sigma_h^2. \end{aligned} \tag{9.4.30}$$

Resultado 9.4.5. *A nivel poblacional, β_h es estimado por*

$$B_h = \frac{\sum_{U_h} y_k}{\sum_{U_h} y_k} \qquad \text{para todo } h = 1, \ldots, H. \tag{9.4.31}$$

A nivel muestral, B_h es estimado por

$$\hat{B}_h = \frac{\sum_{S_h} y_k}{\sum_{S_h} x_k} \qquad \text{para todo } h = 1, \ldots, H. \tag{9.4.32}$$

Resultado 9.4.6. *Bajo un diseño aleatorio simple el estimador de razón separada bajo el modelo (9.4.30), su varianza aproximada y la estimación de la varianza están dados por*

$$\hat{t}_{y,rs} = \sum_{h=1}^{H} t_{xh} \frac{\sum_{S_h} y_k}{\sum_{S_h} x_k} \tag{9.4.33}$$

$$Var_{MAS}(\hat{t}_{y,rs}) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right)\sum_{h=1}^{H} W_h S^2_{E_{U_h}} \tag{9.4.34}$$

$$\widehat{Var}_{MAS}(\hat{t}_{y,rs}) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right)\sum_{h=1}^{H} w_h S^2_{w_{S_h}} \tag{9.4.35}$$

respectivamente. Donde $W_h = (N_h - 1)/(N - 1)$, $w_h = (n_h - 1)/(n - 1)$, $S^2_{E_{U_h}}$ *es la varianza poblacional de la variable linealizada* E_k *en el estrato* h*-ésimo y* $S^2_{e_{S_h}}$ *es la varianza muestral de los valores* e_k *en el estrato* h*-ésimo. Nótese que*

$$E_k = y_k - \left(\frac{\sum_{U_h} y_k}{\sum_{U_h} y_k}\right) x_k \qquad \text{para todo } k \in U_h, \quad h = 1, \ldots, H$$

$$e_k = y_k - \left(\frac{\sum_{S_h} y_k}{\sum_{S_h} y_k}\right) x_k \qquad \text{para todo } k \in U_h, \quad h = 1, \ldots, H$$

Prueba. El espíritu del estimador está enfocado en que como para cada subgrupo se tiene un modelo de razón diferente, entonces se deben sumar todos los estimadores resultantes para cada estrato y así se obtiene una estimación insesgada del total poblacional de la característica de interés, en otras palabras el estimador de t_y está dado por

$$\hat{t}_{y,rs} = \sum_{h=1}^{H} \hat{t}_{yh,rs} = \sum_{h=1}^{H} t_{xh} \frac{\sum_{S_h} y_k}{\sum_{S_h} x_k}$$

donde $\hat{t}_{yh,rs}$ hace referencia al estimador de razón en el h-ésimo estrato. Por otra parte para probar la aproximación de la varianza se debe tener en cuenta que tanto el promedio poblacional como el promedio en cada estrato de las variables linealizadas E_k es nulo. Es decir $\bar{E}_U = \bar{E}_{U_h} = 0$ puesto que

$$\begin{aligned}
\sum_U E_k &= \sum_{h=1}^{H} \sum_{k \in U_h} E_k \\
&= \sum_{h=1}^{H} \sum_{k \in U_h} (y_k - B_h x_k) \\
&= \sum_{h=1}^{H} \sum_{k \in U_h} \left(y_k - \frac{\bar{y}_{U_h}}{\bar{x}_{U_h}} x_k\right) \\
&= \sum_{h=1}^{H} \left(\sum_{k \in U_h} y_k - \frac{\sum_{k \in U_h} y_k}{\sum_{k \in U_h} x_k} \sum_{k \in U_h} x_k\right) \\
&= \sum_{h=1}^{H} \left(\sum_{k \in U_h} y_k - \sum_{k \in U_h} y_k\right) = 0
\end{aligned}$$

Con base en lo anterior se tiene que

$$\begin{aligned}
S^2_{E_U} = \frac{1}{N-1}\sum_{k\in U}(E_k - \bar{E}_U)^2 &= \frac{1}{N-1}\sum_{k\in U}(E_k)^2 \\
&= \frac{1}{N-1}\sum_{h=1}^{H}\sum_{k\in U_h}(E_k)^2 \\
&= \sum_{h=1}^{H}\frac{1}{N-1}\sum_{k\in U_h}(E_k - \bar{E}_{U_h})^2 \\
&= \sum_{h=1}^{H}\frac{N_h-1}{N-1}S^2_{E_{U_h}} = \sum_{h=1}^{H}W_h S^2_{E_{U_h}}
\end{aligned}$$

Al utilizar un razonamiento similar se llega a la demostración de la estimación de la varianza de forma inmediata. ■

En general, para cualquier configuración de diseño de muestreo, cuando se trata de modelos que involucren grupos poblacionales se habla de modelos combinados o separados. Cuando el modelo es combinado, se suponen dos cosas: la primera es que todo el modelo aplica a todas las unidades de la población y eso se ve reflejado directamente en la forma del estimador y la segunda es que se supone que la estrategia de muestreo está inducida por un diseño de muestreo estratificado. Por otro lado cuando el modelo es separado, sólo se supone que la población está separada en subgrupos poblacionales y que existe un modelo distinto para subgrupo. Sea cual sea la configuración en el diseño de muestreo se debe tener en cuenta que para cada grupo existe un modelo. Por último, así como en esta sección se habló del estimador de razón separada o combinada, también es posible hablar del estimador de regresión simple separado o combinado o aún más general, del estimador de regresión general, combinado o separado. En resumen, el adjetivo combinado implica la existencia de estratos inducidos directamente por la población, mientras que el adjetivo separado implica la existencia de subgrupos poblacionales inducidos por el modelo.

9.5 Estimador de regresión simple

El estimador de razón encuentra su mejor uso cuando el gráfico de dispersión de los datos describe una línea recta que pasa por el origen y si la dispersión de la característica de interés aumenta proporcionalmente a la característica de información auxiliar. Sin embargo, es común que el gráfico de dispersión de los datos describa una línea recta que no pasa por el origen o que la dispersión de la característica de interés sea invariante a la característica de información auxiliar. En este último caso, se debe utilizar un estimador de regresión simple pues el modelo que lo sustenta tiene en cuenta las características mencionadas anteriormente.

El estimador de regresión simple está motivado por el **modelo de regresión simple con intercepto** que supone la existencia de una sola variable de información auxiliar la cual presenta una fuerte relación lineal con la característica de interés.

La característica que identifica a este modelo es que la recta de regresión ajustada no pasa por el origen sino que presenta un intercepto significativo. Särndal, Swensson & Wretman (1992) afirman que un modelo que contenga el término del intercepto dará un mejor estimador de regresión que el modelo de media común o el modelo de razón. De esta manera $p = 2$, $\mathbf{x}_k = (1, x_k)'$ y $c_k = 1$ para todo $k \in U$. La formulación del modelo está dada por

$$
\begin{aligned}
Y_k &= \mathbf{X}_k'\boldsymbol{\beta} + \varepsilon_k \\
Y_k &= \beta_0 + \beta_1 X_k + \varepsilon_k
\end{aligned}
\tag{9.5.1}
$$

Donde cada un de los ε_k, $k \in U$, son variables aleatorias independientes e idénticamente distribuidas con media cero y varianza σ^2. Para este modelo $\boldsymbol{\beta}' = (\beta_0, \beta_1)$. Como resultado de lo anterior se tiene que

$$
\begin{aligned}
E_\xi(Y_k) &= \mathbf{x}_k'\boldsymbol{\beta} = \beta_0 + \beta_1 x_k \\
Var_\xi(Y_k) &= \sigma^2.
\end{aligned}
\tag{9.5.2}
$$

Cuando se tiene información auxiliar de tipo continuo que presenta una fuerte relación lineal con la característica de interés y describe un intercepto alejado del origen es común utilizar el estimador de regresión simple. Una de las características del modelo de regresión simple es que la varianza se supone constante en toda la población. Es decir, a medida que aumentan los valores de la característica de información auxiliar la dispersión en la característica de interés se mantiene igual. El siguiente gráfico presenta un ejemplo del tipo de situaciones en las que es pertinente utilizar un estimador de regresión simple.

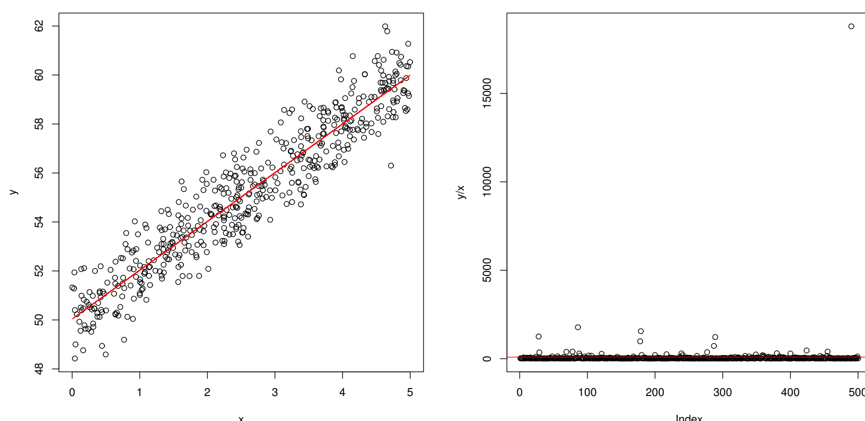


Figura 9.3: *Relación en un modelo de regresión simple con intercepto.*

Si se tuviese acceso a toda la población finita, el estimador del coeficiente de regresión β estaría dado por la minimización de la siguiente función de dispersión

$$D = \sum_U \frac{(y_k - b_0 - b_1 x_k)^2}{\sigma^2}. \tag{9.5.3}$$

Utilizando el resultado 8.4.2 y recurriendo a la ecuación (8.4.6), el estimador B en la población finita toma la siguiente forma

$$\mathbf{B} = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} \tag{9.5.4}$$

En donde

$$b_1 = \frac{\sum_U (x_k - \bar{x}_U)(y_k - \bar{y}_U)}{\sum_U (x_k - \bar{x}_U)^2} \tag{9.5.5}$$

y

$$b_0 = \bar{y}_U - b_1 \bar{x}_U \tag{9.5.6}$$

Por supuesto, como en la práctica sólo se tiene acceso a una muestra particular de población finita, B debe ser estimado de tal manera que siguiendo el resultado 8.4.3. el estimador del coeficiente de regresión basado en la muestra está dado por

$$\hat{\mathbf{B}} = \begin{pmatrix} \hat{b}_0 \\ \hat{b}_1 \end{pmatrix} \tag{9.5.7}$$

En donde

$$\hat{b}_1 = \frac{\sum_S \frac{(x_k - \widetilde{x}_S)(y_k - \widetilde{y}_S)}{\pi_k}}{\sum_S \frac{(x_k - \widetilde{x}_S)^2}{\pi_k}} \tag{9.5.8}$$

y

$$\hat{b}_0 = \widetilde{y}_S - \hat{b}_1 \widetilde{x}_S \tag{9.5.9}$$

Con estas herramientas es posible ahora construir un estimador del total poblacional de la característica de interés el cual está dado por el siguiente resultado.

Resultado 9.5.1. *Bajo el modelo de regresión simple, el estimador del total poblacional está dado por*

$$\hat{t}_{y,rs} = N \left[\widetilde{y}_S + \hat{b}_1 (\bar{x}_U - \widetilde{x}_S) \right] \tag{9.5.10}$$

cuya varianza aproximada es

$$AVar(\hat{t}_{y,rs}) = \sum\sum_U \Delta_{kl} \frac{E_k}{\pi_k} \frac{E_l}{\pi_l}. \tag{9.5.11}$$

con

$$E_k = y_k - \bar{y}_U - b_1 (x_k - \bar{x}_U). \tag{9.5.12}$$

El estimador de la varianza es

$$\widehat{Var}(\hat{t}_{y,rs}) = \sum\sum_S \frac{\Delta_{kl}}{\pi_{kl}} \frac{e_k}{\pi_k} \frac{e_l}{\pi_l} \tag{9.5.13}$$

con

$$e_k = y_k - \widetilde{y}_S - \hat{b}_1(x_k - \widetilde{x}_S). \tag{9.5.14}$$

Prueba. En primer lugar considere el vector de totales de información auxiliar y el vector de estimaciones de los totales de la información auxiliar dado por

$$\mathbf{t_x} = (N, t_x)' \qquad \text{y por} \qquad \hat{\mathbf{t}}_{\mathbf{x},\pi} = (\hat{N}_\pi, t_{x,\pi})'$$

respectivamente. Ahora, recurriendo a la expresión general (9.2.11) y teniendo en cuenta que $\hat{t}_{y,\pi} = \hat{N}_\pi \widetilde{y}_S$, se tiene que

$$\begin{aligned}
\hat{t}_{y,rs} &= \hat{t}_{y,\pi} + (\mathbf{t_x} - \hat{\mathbf{t}}_{\mathbf{x},\pi})'\hat{\mathbf{B}} \\
&= \hat{t}_{y,\pi} + \hat{b}_0(N - \hat{N}_\pi) + \hat{b}_1(t_x - t_{x,\pi}) \\
&= \hat{t}_{y,\pi} + (\widetilde{y}_S - \hat{b}_1\widetilde{x}_S)(N - \hat{N}_\pi) + \hat{b}_1(t_x - t_{x,\pi}) \\
&= N\widetilde{y}_S - N\hat{b}_1\widetilde{x}_S + N\hat{b}_1 t_x \\
&= N\left[\widetilde{y}_S + \hat{b}_1(\bar{x}_U - \widetilde{x}_S)\right]
\end{aligned}$$

Por otra lado, el cálculo de la varianza aproximada y la estimación de la varianza del estimador de razón se obtienen al utilizar el resultado 9.2.3. y considerando el producto vectorial

$$\begin{aligned}
\mathbf{x}_k'\mathbf{B} &= (1, x_k)\begin{pmatrix} b_0 \\ b_1 \end{pmatrix} \\
&= b_0 + b_1 x_k
\end{aligned}$$

Por tanto

$$\begin{aligned}
E_k &= y_k - \mathbf{x}_k'\mathbf{B} \\
&= y_k - (b_0 + b_1 x_k) \\
&= y_k - (\bar{y}_U - b_1\bar{x}_U + b_1 x_k) \\
&= y_k - \bar{y}_U - b_1(x_k - \bar{x}_U)
\end{aligned}$$

Análogamente para el caso de e_k. ■

Nótese que fácilmente se puede demostrar que

$$\sum_s \frac{e_k}{\pi_k} = 0$$

y, por lo tanto, es posible escribir el estimador de regresión simple en una forma simplificada

$$\hat{t}_{y,rs} = \sum_U \hat{y}_k \tag{9.5.15}$$

$$= \sum_U \mathbf{x}'_k \hat{\mathbf{B}} \tag{9.5.16}$$

Apelando a las expresiones (9.2.16) y (9.2.17) se tiene que

$$g_{ks} = \frac{N}{\hat{N}_\pi}\left[1 + a_S(x_k - \widetilde{x}_S)\right] \tag{9.5.17}$$

Con

$$a_S = \frac{\bar{x}_U - \widetilde{x}_S}{\widetilde{S}^2_{x_S}}$$

y

$$\widetilde{S}^2_{x_S} = \frac{\sum_S x_k - \widetilde{x}_S/\pi_k}{\sum_S 1/\pi_k}$$

Särndal, Swensson & Wretman (1992) manifiestan que en algunas encuestas es preferible utilizar el estimador de razón sobre el estimador de regresión simple, incluso cuando este último tiene grandes ventajas en la eficiencia sobre el estimador de razón. Uno de los motivos de esta preferencia es que el estimador de razón tiene una estructura muy simple que resalta cuando se necesitan estimaciones tanto del total poblacional de la característica de interés, como de la razón entre dos totales.

Aunque el estimador de regresión es más eficiente que el estimador de razón cuando la línea de regresión no pasa por el origen, es importante notar que esta eficiencia resalta mucho más cuando la población y la muestra consta de un número suficientemente grande de individuos. Esto quiere decir que cuando la encuesta es pequeña o mediana, estas propiedades pueden no cumplirse con rigurosidad.

9.5.1 Marco y Lucy

Retomando la población de empresas pertenecientes al sector industrial, suponga que se desea estimar el total de las características de interés Ingreso y Número de Empleados mediante un estimador de regresión simple con intercepto que obedezca al modelo dado por la expresión (9.5.2), en donde las características de interés están relacionadas con la característica de información auxiliar Impuestos (declarados en el último año fiscal) y que supone el mismo comportamiento estructural a lo largo de toda la población. Suponga que se selecciona una muestra aleatoria simple de tamaño $n = 400$.

```
> data(Lucy)
> attach(Lucy)
> N <- dim(Lucy)[1]
> n <- 400
```

```
> sam <- S.SI(N,n)
> muestra <- Lucy[sam,]
> attach(muestra)
```

Para obtener la estimación del total de las características de interés se define la matriz de información auxiliar, que en este caso particular corresponde a una matriz conteniendo dos vectores columna: el primero, un vector de unos y el segundo un vector de valores correspondientes a la observación de los impuestos declarados por las empresas seleccionadas en la muestra aleatoria. El siguiente paso es la utilización de la función `GREG.SI` del paquete `TeachingSampling` definiendo correctamente los parámetros según el modelo dado por (9.5.2)

```
> estima <- data.frame(Income, Employees)
> pik <- rep(n/N,n)
> x <- Taxes
> tx <- c(N,28654)
> b <- E.Beta(estima,x,pik,ck=1,b0=TRUE)
> GREG.SI(N,n,estima,x,tx, b, b0=TRUE)
                                Income    Employees
Total estimado            1.030302e+06 1.506098e+05
Varianza                  1.534930e+08 7.299305e+06
coeficiente de variación 1.202485e+00 1.793856e+00
```

El siguiente código muestra que la propiedad de calibración se mantiene para este modelo. Véase que la multiplicación de los valores observados de cada una de las características de interés con los pesos de calibración da como resultado la misma estimación arrojada por el anterior código.

```
> w <- Wk(x,tx,pik,ck=1,b0=TRUE)
> sum(x*w)
[1] 28654
> sum(1*w)
[1] 2396
```

A continuación se muestra la tabla de resultados que indica que asumiendo este modelo se gana mucha más precisión y eficiencia que en el caso en donde no se asume ningún modelo o el modelo asumido no es acorde con el estado de la naturaleza del sector industrial.

Tabla 9.3: *modelo de regresión simple: estimación de los totales de las características de interés.*

Variable	Total poblacional	Total estimado	cve %	Desv. %
Ingreso	1035217	1030302	1.20	0.47
Empleados	151950	150609	1.79	0.88

9.6 Post-estratificación

En muchas situaciones prácticas es útil estratificar la población. Sin embargo, si el marco de muestreo no incluye la variable de estratificación, para todas las unidades de la población, es imposible pensar en un diseño de muestreo estratificado[3]. Cochran (1977) afirma que algunos ejemplos particulares en donde no se conoce la pertenencia de las unidades antes de la recolección de los datos son características personales como la edad, el género, la raza y el nivel educativo. Nótese que el total poblacional de las anteriores variables, que inducen subgrupos poblacionales, puede ser conocido recurriendo a alguna fuente externa confiable o a estadísticas oficiales. Por lo tanto, es posible construir un estimador que tenga en cuenta el conocimiento de estos totales y que incorpore la información recolectada en la muestra acerca de la pertenencia de cada unidad a un subgrupo poblacional (Holt & Smith 1979, Jagers 1986).

Antes de continuar con la construcción de dicho estimador, es pertinente hacer un recuento de cómo influyen los subgrupos poblacionales en el desarrollo de la teoría de muestreo. De esta manera, el estadístico estará en capacidad de discernir acerca de la mejor estrategia de muestreo cuando su investigación involucre la estimación en subgrupos poblacionales.

9.6.1 Subgrupos poblacionales

Aunque el marco de referencia de la teoría de muestreo es la estimación de un parámetro de interés sobre alguna característica de interés, lo cierto es que en la práctica no sólo se necesitan estimaciones que cobijen la población entera sino que también son indispensable estimaciones que involucren subgrupos poblacionales.

Los subgrupos poblacionales inducen una partición de la población definida de la siguiente manera. Sean $U_1, \ldots, U_g, \ldots, U_G$ subgrupos poblacionales tales que

$$\bigcup_{g=1}^{G} U_g = U; \qquad U_g \bigcap U_h = \phi, \quad \text{si } g \neq h \tag{9.6.1}$$

Además si N_g es el tamaño absoluto de U_g se tiene que $\sum_{g=1}^{G} N_g = N$. Por otra parte, los subgrupos poblacionales también inducen una partición en la muestra realizada $S = s$ mediante subconjuntos $s_1, \ldots, s_g, \ldots, s_G$, con $s_g = U_g \bigcap s$, tales que

$$\bigcup_{g=1}^{G} s_g = s; \qquad s_g \bigcap s_h = \phi, \quad \text{si } g \neq h. \tag{9.6.2}$$

Si el tamaño de la muestra realizada es n y el tamaño de s_g es n_g se tiene que $\sum_{g=1}^{G} n_g = n$.

[3]Nótese que un requisito indispensable para realizar un diseño de muestreo estratificado es el conocimiento a priori de la pertenencia de todos y cada uno de las unidades que conforman la población a los estratos propuestos.

En general, cuando se habla de subgrupos poblacionales se está haciendo referencia a **dominios de interés**[4], **estratos**[5] o **post-estratos**[6]. Cuando se hace referencia a subgrupos poblacionales, es indispensable conocer en qué se diferencian cada uno de ellos pues de esto depende que las investigaciones arrojen resultados confiables mediante el planteamiento de la mejor estrategia de muestreo.

- **Dominios de interés:** este tipo de subgrupos poblacionales son aquellos para los cuales *se requieren estimaciones separadas* del parámetro de interés. Estos requerimientos se planean en la etapa de diseño para asegurar que el diseño de la muestra sea tal que al momento de la recolección de la información exista una buena cobertura en cada uno de los dominios de interés. Lo anterior sólo se puede lograr ampliando el tamaño de muestra n puesto que *el marco de muestreo no informa acerca de la pertenencia de los individuos* a los dominios de interés. Los aspectos más importante de esta clase de subgrupos poblacionales son que *el número de individuos en la muestra que pertenecen a un dominio n_d de interés es siempre aleatorio*, y para algunos dominios particulares puede llegar a ser muy pequeño. Por otro lado, el *tamaño absoluto de cada dominio N_d no se conoce* ni antes de la etapa de diseño ni después de la etapa de estimación.

- **Estratos:** cuando *el marco de muestreo permite conocer la pertenencia de todos los individuos* de la población a un subgrupo poblacional, se dice que esta clase de subgrupos se llaman estratos. Más aun, cuando se sabe que la característica de interés tiene un comportamiento distinto en cada uno de los estratos y se planea un diseño de muestreo que tenga en cuenta este aspecto mediante la selección aleatoria de unidades en cada uno de los estratos, se dice que el diseño de muestreo es estratificado. El aspecto fundamental de esta clase de subgrupos poblacionales es que el conocimiento de la pertenencia de los individuos a los estratos se incorpora en la etapa de diseño de la muestra. Nótese que a diferencia de los dominios, en los estratos se conoce tanto N_h como n_h antes de la etapa de estimación.

- **Post-estratos:** la propiedad que caracteriza a este tipo de subgrupos poblacionales es que aunque *en la etapa de diseño el tamaño del post-estrato N_g es conocido, se desconoce el número de individuos que pertenecerán al post-estrato n_g* en la muestra realizada. Al respecto Särndal, Swensson & Wretman (1992) proclaman que existen dos situaciones en las cuales se presenta esta situación, llamada post-estratificación:

  1. El marco de muestreo es tal que se conoce la pertenencia de todos los elementos a los subgrupos poblacionales pero el investigador decide no utilizar esta información en la etapa de diseño. Las razones para esto son diversas pero principalmente se decide obviar este tipo de información por practicidad logística. Una vez que se ha realizado la selección de la muestra, se observa la característica de interés y_k en los individuos tales que $k \in S$. El investigador decide utilizar la información auxiliar de pertenencia a los post-estratos en la etapa de estimación para mejorar

[4] La notación para los dominios un dominio de interés es U_d, $d = 1, \ldots, D$.
[5] La notación para los estratos es U_h, $h = 1, \ldots, H$.
[6] La notación para los post-estratos es U_g, $g = 1, \ldots, G$.

la eficiencia de la estrategia de muestreo, en particular del estimador propuesto.

2. Mediante alguna fuente de información confiable[7] se conocen los tamaños absolutos N_g de cada subgrupo poblacional aunque se desconoce la pertenencia de los individuos a los subgrupos pues el marco de muestreo presenta esta deficiencia. Después de la etapa de diseño, se observa la característica de interés y se pregunta acerca de la pertenencia de los individuos seleccionados en los post-estratos de tal forma que en la etapa de estimación se utiliza esta información para mejorar la eficiencia de los estimadores de los parámetros de interés[8].

Es en el caso de los subgrupos poblacionales llamados post-estratos que nos enfocaremos en las próximas secciones, describiendo cómo es posible crear estimadores eficientes bajo este marco de referencia y cómo influye un modelo de análisis de varianza (ANOVA) en el desarrollo teórico desde el punto de vista de los modelos lineales.

9.6.2 El estimador de post-estratificación

El estimador de post-estratificación utiliza la incorporación de la información auxiliar[9] para reducir el sesgo generado por el desconocimiento de la pertenencia de los individuos a los subgrupos poblacionales. En general, el uso de la información auxiliar en la etapa de estimación mejora la calidad de la encuesta puesto que (Zhang 2000):

1. Reduce la varianza muestral de las características altamente relacionadas con la información auxiliar.

2. Reduce el sesgo, especialmente el sesgo causado por errores no muestrales tales como la ausencia de respuesta y la subcobertura del marco de muestreo.

3. Impone consistencia con los resultados obtenidos mediante fuentes oficiales.

El estimador de post-estratificación tiene muchas ventajas; entre ellas es el estimador máximo verosímil bajo un marco de referencia muy general en el cual se demuestra que tiene varianza mínima (Jagers 1986) por tanto este estimador ofrece protección contra configuraciones muestrales desfavorables siendo considerado como un estimador robusto (Holt & Smith 1979).

El estimador de post-estratificación puede ser visto como un caso particular del estimador general de regresión donde las características de información auxiliar corresponden a variables indicadoras de los subgrupos poblacionales. Este enfoque fue adoptado por Bethelehem & Keller (1987) y en esta sección se le dará el mismo enfoque. En general, el proceso de post-estratificación se desarrolla bajo dos modelos grupales que inducen diferentes estimadores conocidos como el estimador

[7] Censos, registros externos o estadísticas oficiales

[8] En este apartado el investigador debe planear muy bien el diseño muestral de tal manera que exista una buena cobertura en los post-estratos puesto que el estimador puede resultar sesgado.

[9] Mediante los totales poblacionales N_g de los post-estratos

de media post-estratificada y el estimador de razón post-estratificada. Los siguientes apartados darán cuenta de la concepción y construcción de cada uno de ellos desde el punto de vista de los modelos lineales.

9.6.3 Estimador de media post-estratificada

La incorporación de la información auxiliar en el estimador de post-estratificación se hizo mucho antes de que se utilizaran los modelos lineales en la construcción de un estimador asistido por modelos. Sin embargo, siguiendo el hilo conductor del capítulo, realizaremos la construcción del mismo estimador de post-estratificación mediante la incorporación de un modelo lineal de tipo ANOVA (o de media post-estratificada) que supone la existencia de una sola característica auxiliar de tipo discreto[10] que tiene G niveles o post-estratos. Un supuesto fuerte alrededor del modelo ANOVA es que la característica de interés es altamente correlacionada con los post-estratos inducidos por esta característica discreta. Se tiene que $p = G$, $\mathbf{x}_k = \mathbf{d}_k = (\underbrace{0, 0, \ldots, 1, \ldots, 0, 0}_{G \texttt{ grupos}})'$ y $c_k = 1$ para todo $k \in U$. De esta manera, el vector de totales de la característica de información auxiliar discreta está dado por

$$\mathbf{t_x} = (N_1, \ldots, N_g, \ldots, N_g)' \tag{9.6.3}$$

y el vector de totales estimados de la característica de información auxiliar discreta está dado por

$$\hat{\mathbf{t}}_{\mathbf{x},\pi} = (\hat{N}_{1,\pi}, \ldots, \hat{N}_{g,\pi}, \ldots, \hat{N}_{g,\pi})' \tag{9.6.4}$$

La formulación del modelo está dada por

$$Y_k = \mathbf{d}_k'\boldsymbol{\beta} + \varepsilon_k = \beta_g + \varepsilon_k \tag{9.6.5}$$

Donde $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_g, \ldots, \beta_G)'$ y cada uno de los ε_k $k \in U$ son variables aleatorias independientes e idénticamente distribuidas con media cero y varianza σ_g^2. Nótese que $\mathbf{d}_k = (d_{1k}, \ldots, d_{gk}, \ldots, d_{Gk})'$ con

$$d_{gk} = \begin{cases} 1, & \text{si } k \in U_g \\ 0, & \text{en otro caso.} \end{cases} \tag{9.6.6}$$

Luego el modelo de super-población está dado por

$$\begin{aligned} E_\xi(Y_k) &= \mathbf{d}_k'\boldsymbol{\beta} = \beta_g \\ Var_\xi(Y_k) &= \sigma_g^2. \end{aligned} \tag{9.6.7}$$

Särndal, Swensson & Wretman (1992) explican que este modelo es eficiente cuando la característica de interés es homogénea dentro de los post-estratos U_g ($g =$

[10] A diferencia de los demás estimadores considerados en este capítulo que suponen la existencia de una característica auxiliar de tipo continuo.

$1, \ldots, G$) pero disímil y heterogénea entre cada uno de los post-estratos. Cuando las anteriores condiciones se satisfacen entonces el modelo ANOVA explicará una gran parte de la dispersión de la característica de interés. Es común utilizar el estimador de post-estratificación cuando el gráfico de dispersión entre la característica de interés y la característica de información auxiliar define una recta de regresión constante y paralela al eje de las abscisas, para todos y cada uno de los post-estratos, como se puede observar la figura 9.4.

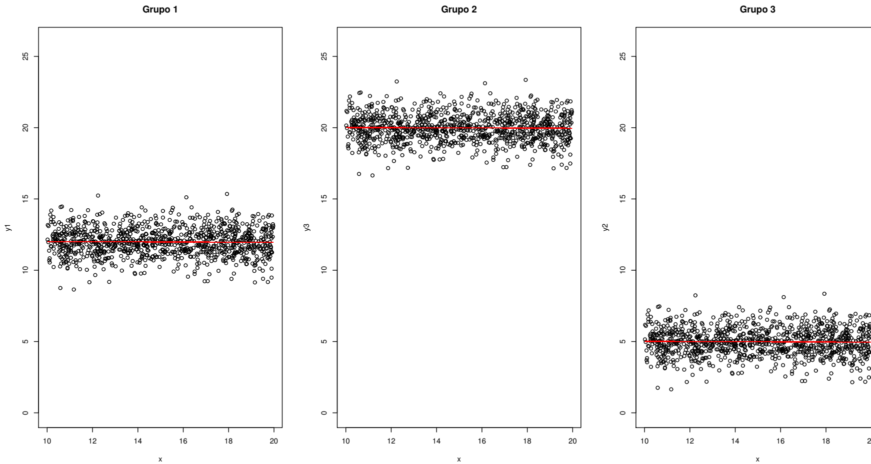


Figura 9.4: *Relación en un modelo de media post-estratificada.*

Si se tuviese acceso a toda la población finita, el estimador del coeficiente de regresión $\boldsymbol{\beta}$ estaría dado por la minimización de la siguiente función de dispersión

$$D = \sum_{U} \frac{(y_k - \mathbf{x}_k' \mathbf{B})^2}{\sigma^2}. \qquad (9.6.8)$$

Utilizando el resultado 8.4.2 y recurriendo a la ecuación (8.4.6), el estimador $\mathbf{B}$ en la población finita toma la siguiente forma

$$\mathbf{B} = (B_1, B_2, \ldots, B_G)' \qquad (9.6.9)$$

En donde

$$B_g = \frac{\sum_{U_g} y_k}{N_g} \qquad g = 1, \ldots, G \qquad (9.6.10)$$

Por supuesto, como en la práctica sólo se tiene acceso a una muestra particular de población finita, $\mathbf{B}$ debe ser estimado de tal manera que siguiendo el resultado 8.4.3. llegamos a la siguiente expresión

$$\hat{\mathbf{B}} = (\hat{B}_1, \hat{B}_2, \ldots, \hat{B}_G)' \qquad (9.6.11)$$

En donde

$$\hat{B}_g = \left(\sum_{S_g} \frac{1}{\pi_k}\right)^{-1} \left(\sum_{S_g} \frac{y_k}{\pi_k}\right) = \frac{\hat{t}_{yU_g,\pi}}{\hat{N}_{U_g,\pi}} = \widetilde{y}_{S_g} \quad (9.6.12)$$

Con estas herramientas es posible ahora construir un estimador del total poblacional de la característica de interés el cual está dado por el siguiente resultado.

Resultado 9.6.1. *Bajo el modelo de media post-estratificada, el estimador del total poblacional está dado por*

$$\hat{t}_{y,mpos} = \sum_{\substack{g=1 \\ n_h \neq 0}}^{G} N_g \widetilde{y}_{S_g} \quad (9.6.13)$$

$$= \sum_{\substack{g=1 \\ n_h \neq 0}}^{G} \sum_{k \in S_g} \frac{N_g}{\hat{N}_{g,\pi}} \frac{y_k}{\pi_k} \quad (9.6.14)$$

cuya varianza aproximada es

$$AVar(\hat{t}_{y,mpos}) = \sum\sum_{U} \Delta_{kl} \frac{E_k}{\pi_k} \frac{E_l}{\pi_l}. \quad (9.6.15)$$

con

$$E_k = y_k - \bar{y}_{U_g} \qquad g = 1, \ldots, G. \quad (9.6.16)$$

El estimador de la varianza es

$$\widehat{Var}(\hat{t}_{y,mpos}) = \sum\sum_{S} \frac{\Delta_{kl}}{\pi_{kl}} \frac{e_k}{\pi_k} \frac{e_l}{\pi_l} \quad (9.6.17)$$

con

$$e_k = y_k - \widetilde{y}_{S_g} \qquad g = 1, \ldots, G. \quad (9.6.18)$$

Prueba. Nótese que el estimador de post-estratificación es un caso particular del estimador general de regresión. Entonces, por la configuración particular de $\mathbf{x}_k$ y adecuando la expresión (9.2.11) se tiene que

$$\begin{aligned}
\hat{t}_{y,mpos} &= \hat{t}_{y,\pi} + (\mathbf{t_x} - \hat{\mathbf{t}}_{\mathbf{x},\pi})'\hat{\mathbf{B}} \\
&= \hat{t}_{y,\pi} + (N_1 - \hat{N}_{1,\pi})\widetilde{y}_{S_1} + \ldots + (N_G - \hat{N}_{G,\pi})\widetilde{y}_{S,G} \\
&= N_1\widetilde{y}_{S_1} + \ldots + N_G\widetilde{y}_{S_G} \\
&= \sum_{g=1}^{G} N_g \widetilde{y}_{S_g}
\end{aligned}$$

El cálculo de la varianza aproximada y la estimación de la varianza del estimador de razón son inmediatos al utilizar el resultado 9.2.3. Notando que

$$E_k = y_k - \mathbf{x}_k'\mathbf{B} = y_k - B_g = y_k - \bar{y}_{U_g}$$

Análogamente con e_k. ■

De manera similar al estimador de media común, el espíritu de este estimador está en la corrección que hace al estimador de Horvitz-Thompson mediante el cociente $\frac{N_g}{\hat{N}_{g,\pi}}$. De esta forma, cuando el estimador de Horvitz-Thompson está subestimando o sobreestimando el total poblacional, entonces este cociente corrige inmediatamente esta sub o sobre estimación para cada uno de los post-estratos o subgrupos poblacionales.

Por otro lado, fácilmente se puede demostrar que

$$\sum_s \frac{e_k}{\pi_k} = 0$$

Lo anterior se tiene puesto que,

$$\begin{aligned}\sum_S \frac{e_k}{\pi_k} &= \sum_S \frac{y_k}{\pi_k} - \sum_S \frac{\mathbf{d}_k{}'}{\pi_k}\hat{\mathbf{B}}\\ &= t_{y,\pi} - \hat{N}_{U_1,\pi}\frac{\hat{t}_{y_{U_1},\pi}}{\hat{N}_{U_1,\pi}} - \ldots - \hat{N}_{U_1,\pi}\frac{\hat{t}_{y_{U_G},\pi}}{\hat{N}_{U_1,\pi}} \quad = t_{y,\pi} - \sum_{g=1}^{G}\hat{t}_{y_{U_g},\pi} = 0\end{aligned}$$

Como consecuencia de esto, es posible escribir al estimador de media común en la forma simplificada, teniendo en cuenta que $\hat{y}_k = \mathbf{x}_k'\hat{\mathbf{B}} = \hat{B}_g = \widetilde{y}_{S_g}$ cuando $k \in U$. Por tanto

$$\hat{t}_{y,mpos} = \sum_U \hat{y}_k = \sum_{g=1}^{G}\sum_{U_g}\hat{y}_k = \sum_{g=1}^{G} N_g \widetilde{y}_{S_g}$$

Además recurriendo a las expresiones (9.2.16) y (9.2.17) se tiene que

$$g_{ks} = 1 + \frac{N_g - \hat{N}_{g,\pi}}{\hat{N}_{g,\pi}} = \frac{N_g}{\hat{N}_{g,\pi}} \tag{9.6.19}$$

Al respecto de este estimador de media post-estratificada (Lohr 2000) explica que es posible obtener varianzas muy pequeñas cuando los post-estratos se conforman después de la recolección de la información así como es posible obtener diferentes p-valores cuando se decide acerca de la hipótesis nula después de la observación de los datos.

Diseño de muestreo aleatorio simple

Es en este diseño de muestreo que el estimador de media post-estratificación toma su expresión más utilizada y más conocida. De hecho, en la jerga técnica estadística cuando se menciona el **estimador de post-estratificación**, se está haciendo referencia al estimador de media post-estratificada bajo un diseño de muestreo aleatorio simple.

Resultado 9.6.2. *Si el diseño de muestreo es aleatorio simple, el estimador de media post-estratificada, su varianza aproximada y el estimador de la varianza están dados por las siguientes expresiones*

$$\hat{t}_{y,mpos} = \sum_{\substack{g=1\\ n_h\neq 0}}^{G} N_g \bar{y}_{S_g} \tag{9.6.20}$$

$$Var_{MAS}(\hat{t}_{y,mpos}) = \frac{N^2}{n}\left(1-\frac{n}{N}\right)\sum_{g=1}^{G}\frac{N_g-1}{N-1}S^2_{E_{U_g}} \tag{9.6.21}$$

$$\widehat{Var}_{MAS}(\hat{t}_{y,mc}) = \frac{N^2}{n}\left(1-\frac{n}{N}\right)\sum_{g=1}^{G}\frac{n_g-1}{n-1}S^2_{e_{S_g}} \tag{9.6.22}$$

respectivamente. Donde $\bar{y}_{U_g} = \sum_{k\in U_g} y_k/N_g$ *y* $\bar{y}_{S_g} = \sum_{k\in S_g} y_k/n_g$, $S^2_{E_{U_g}}$ *es la varianza poblacional de los errores* $E_k = y_k - \bar{y}_{U_g}$ *para* $k\in U_g$ *dada por*

$$S^2_{E_{U_g}} = \frac{1}{N_g-1}\sum_{k\in U_g}(y_k-\bar{y}_{U_g}), \tag{9.6.23}$$

y $S^2_{e_{S_g}}$ *es la varianza muestral de los errores* $e_k = y_k - \bar{y}_{S_g}$ *para* $k\in S_g$ *dada por*

$$S^2_{e_{S_g}} = \frac{1}{n_g-1}\sum_{k\in S_g}(y_k-\bar{y}_{S_g}). \tag{9.6.24}$$

Prueba. El resultado se sigue inmediatamente al aplicar los principios del estimador de Horvitz-Thompson a las expresiones (9.3.7) y (9.3.10) bajo el diseño de muestreo aleatorio simple. Nótese que bajo el diseño de muestreo aleatorio simple, $\bar{E}=0$ y $\bar(e)=0$. Por lo tanto, partiendo de

$$\frac{N^2}{n}\left(1-\frac{n}{N}\right)\sum_{k\in U}S^2_{E_U} \tag{9.6.25}$$

y teniendo en cuenta que

$$S^2_{E_U} = \frac{1}{N-1}\sum_{k\in U}(E_k - \underbrace{\bar{E}}_{cero})^2 \tag{9.6.26}$$

$$= \frac{1}{N-1}\sum_{g=1}^{G}\sum_{k\in U-g}(y_k - \bar{y}_{U_g})^2 \tag{9.6.27}$$

$$= \sum_{g=1}^{G}\frac{N_g-1}{N-1}S^2_{E_{U_g}} \tag{9.6.28}$$

se tiene el resultado de la varianza. Se deben seguir pasos análogos para encontrar la expresión de la estimación de la varianza. ■

Nótese que aunque el estimador de post-estratificación toma una expresión muy similar al estimador de Horvitz-Thompson bajo un diseño estratificado. Es en este punto donde Smith (1991) se pregunta cómo evaluar las propiedades del estimador... ¿se debería retornar al diseño aleatorio simple (como se refleja en el resultado anterior) o se debería utilizar las propiedades del diseño estratificado? La respuesta a la anterior pregunta se ha convertido en objeto de una ardua investigación. Sin embargo, desde un punto de vista pragmático, es muy común encontrar que en muchas investigaciones la inferencia que se hace con respecto al estimador de post-estratificación sea condicional al tamaño de muestra de cada uno de los subgrupos y, por consiguiente, las expresiones de varianza no son las mismas que las mostradas en el anterior resultado. Sobre esta discusión, Tille & Ardilly (2006) menciona que este estimador es insesgado siempre y cuando en la muestra realizada existan individuos pertenecientes a todos y cada uno de los post-estratos; es decir $n_g \neq 0$. Ahora, este tipo de análisis está condicionado al evento $n_g(S) = n_g \neq \forall g = 1, \ldots, G$ y los resultados que se presentan a continuación siguen esa idea[11].

En primer lugar nótese que el tamaño de muestra en cada post-estrato n_g, $\quad g = 1, \ldots, G$. es una variable aleatoria y, de esta manera, se tiene la siguiente propiedad.

Resultado 9.6.3. *La esperanza y varianza de la variable aleatoria n_g están dadas por*

$$E(n_g) = n\frac{N_g}{N} \qquad Var(n_g) = n\frac{N_g}{N}\frac{N-N_g}{N}\frac{N-n}{N-1} \qquad g = 1, \ldots, G. \tag{9.6.29}$$

respectivamente. La probabilidad de que al seleccionar una muestra, un post-estrato quede sin representación está dada por

$$Pr(n_g = 0) = \frac{(N-N_g)!}{(N-N_g-n)!}\frac{(N-n)!}{N!}$$

donde $N = \sum_{g=1}^{G} N_h$ y $n = \sum_{g=1}^{G} n_h$.

[11] Nótese que fácilmente se pudieron desarrollar las fórmulas de varianza de manera incondicional partiendo de los resultados que caracterizan al estimador general de regresión.

Prueba. La distribución de n_g es hipergeométrica definida como

$$Pr(n_g = r) = \frac{\binom{N_g}{r}\binom{N-N_g}{n-r}}{\binom{N}{n}} \qquad r = 0, \ldots, n.$$

Al desarrollar la anterior expresión cuando $r = 0$ se llega al resultado. ■

Con el resultado anterior es posible garantizar que la probabilidad de $n_g = 0$ sea muy pequeña al establecer la siguiente condición

$$n\frac{N_g}{N} > 30. \tag{9.6.30}$$

Por otro lado, para realizar el análisis condicional se define el evento A denotado como

$$n_g \geq 1 \qquad g = 1, \ldots, G. \tag{9.6.31}$$

Se puede demostrar la distribución de la muestra aleatoria S *condicionada a una configuración particular y fija* de los tamaños de muestra en cada uno de los post-estratos, por ejemplo $\mathbf{n}_s = (n_1, \ldots, n_G)$, es idéntica a la probabilidad de selección de una muestra estratificada al tratar a cada subgrupo poblacional como un estrato en el que, de manera independiente, se seleccionan n_g individuos de N_g para cada subgrupo U_g; $g = 1, \ldots, G$.

Resultado 9.6.4. *Dada una configuración fija de los tamaños de muestra en cada estrato y bajo el evento A_1, el estimador de media post-estratificada es insesgado para el total poblacional de la característica de interés. Su varianza y el estimador de la varianza se encuentran dadas por las siguientes expresiones*

$$Var_{MAS}(\hat{t}_{y,mpos}|A, \mathbf{n}_s) = \sum_{g=1}^{G} N_g^2\left(\frac{1}{n_g} - \frac{1}{N_g}\right)S^2_{y_{U_g}} \tag{9.6.32}$$

$$\widehat{Var}_{MAS}\left(\hat{t}_{y,mpos}|A, \mathbf{n}_s\right) = \sum_{g=1}^{G} N_g^2\left(\frac{1}{n_g} - \frac{1}{N_g}\right)S^2_{y_{S_g}} \tag{9.6.33}$$

respectivamente. Nótese que $\widehat{Var}_{MAS}\left(\hat{t}_{y,mpos}|A, \mathbf{n}_s\right)$ es un estimador condicionalmente insesgado de $Var_{MAS}\left(\hat{t}_{y,mpos}|A, \mathbf{n}_s\right)$.

Prueba. Nótese que el estimador de media post-estratificada es idéntico al estimador de Horvitz-Thompson bajo un diseño aleatorio estratificado. Por tanto, utilizando el resultado 5.2.3. se tiene que

$$\begin{aligned}
E_{MAS}(\hat{t}_{y,mpos}|A, \mathbf{n}_s) &= \sum_{\substack{g=1\\ n_h\neq 0}}^{G} N_g E_{MAS}(\bar{y}_{S_g})\\
&= \sum_{\substack{g=1\\ n_h\neq 0}}^{G} N_g \bar{y}_{U_g} = t_y
\end{aligned}$$

y una vez más accediendo al resultado 5.2.3 la varianza del estimador está dada por la clásica varianza del diseño aleatorio estratificado, la cual después de un poco de álgebra se convierte en la expresión de este resultado.

$$\begin{aligned} Var_{MAS}(\hat{t}_{y,\pi}\,|A,\mathbf{n}_s) &= \sum_{g=1}^{G}\frac{N_g^2}{n_g}\left(1-\frac{n_g}{N_g}\right)S^2_{y_{U_g}} \\ &= \sum_{g=1}^{G}N_g^2\left(\frac{1}{n_g}-\frac{1}{N_g}\right)S^2_{y_{U_g}} \end{aligned}$$

Análogamente se tiene el resultado para el estimador de la varianza y su insesgamiento viene de que $S^2_{y_{S_g}}$ es condicionalmente insesgado para $S^2_{y_{U_g}}$. ■

Holt & Smith (1979) consideró el problema de la inferencia condicional bajo una muestra dada y concluyen que ésta debería ser usada en la *etapa de estimación* dado que las otras posibles configuraciones distintas a la observada se consideran irrelevantes en el momento de la inferencia. Por otro lado, también es posible realizar un análisis incondicional en la *etapa de diseño* puesto que permite escoger una estrategia de muestreo adecuada.

El análisis incondicional empieza al definir la cantidad $E_1(\cdot)$ como la esperanza bajo la distribución de $\mathbf{n}_s$, la cual es hipergeométrica multivariante. Por tanto se tiene el siguiente resultado.

Resultado 9.6.5. *El estimador de media post-estratificada es incondicionalmente insesgado para el total poblacional de la característica de interés. Su varianza aproximada se encuentran dada por la siguiente expresión*

$$Var_{MAS}(\hat{t}_{y,mpos}\,|A) = \frac{N-n}{n}\sum_{g=1}^{G}N_gS^2_{y_{U_g}} + \frac{N(N-n)}{n^2(n-1)}\sum_{g=1}^{G}(N-N_h)S^2_{y_{U_g}} \quad (9.6.34)$$

Prueba. El insesgamiento del estimador se tiene del resultado 9.6.4 puesto que

$$E_{MAS}(\hat{t}_{y,mpos}\,|A) = E_1(E_{MAS}[\hat{t}_{y,mpos}\,|A,\mathbf{n}_s]) = E_1(t_y) = t_y$$

Por otro lado, nótese que

$$\begin{aligned} Var_{MAS}(\hat{t}_{y,mpos}\,|A) = &\ E_1(Var_{MAS}[\hat{t}_{y,mpos}\,|A,\mathbf{n}_s]) \\ &+ Var_1(E_{MAS}[\hat{t}_{y,mpos}\,|A,\mathbf{n}_s]) \end{aligned}$$

El último sumando de la ecuación anterior es nulo porque

$$Var_1(E_{MAS}[\hat{t}_{y,mpos}\,|A,\mathbf{n}_s]) = Var_1(t_y) = 0$$

y como n_g es una cantidad aleatoria, entonces

$$E_1(Var_{MAS}[\hat{t}_{y,mpos}\,|A,\mathbf{n}_s]) = \sum_{g=1}^{G}N_g^2\left(E\left[\frac{1}{n_g}\right]-\frac{1}{N_g}\right)S^2_{y_{U_g}} \quad (9.6.35)$$

Por tanto se debería calcular $E\left[\frac{1}{n_g}\right]$, pero este proceso es muy dispendioso. En vez de esto, es posible aproximar esta expresión definiendo la siguiente cantidad

$$E\left(\frac{1}{n_g}\right) = \frac{1}{E(n_g)}E\left(\frac{1}{1-\varphi}\right)$$

Donde

$$\varphi = 1 - \frac{n_g}{E(n_g)} = 1 - \frac{N}{n}\frac{N_g}{n_g}$$

Si el tamaño de la muestra se considera grande, es posible afirmar que φ es una cantidad cercana a cero. Por tanto, utilizando un desarrollo de series matemáticas geométricas se tiene que

$$\frac{1}{1-\varphi} = 1 + \varphi + \varphi^2 + \ldots$$

y por consiguiente se obtiene la siguiente aproximación de segundo orden al utilizar la ecuación (9.6.28).

$$\begin{aligned}
E\left(\frac{1}{n_g}\right) &\approx \frac{1}{E(n_g)}E\left(1 + \varphi + \varphi^2\right) \\
&= \frac{N}{nN_g}E\left[1 + \left(1 - \frac{N}{n}\frac{N_g}{n_g}\right) + \left(1 - \frac{N}{n}\frac{N_g}{n_g}\right)^2\right] \\
&= \frac{N}{nN_g}\left[1 + \frac{N^2 Var(n_g)}{n^2 N_g^2}\right] \\
&= \frac{N}{nN_g} + \frac{(N-N_g)N}{N_g^2}\frac{N-n}{n^2(N-1)}
\end{aligned}$$

Por lo tanto reemplazando en (9.6.35) se tiene que

$$Var_{MAS}(\hat{t}_{y,mpos}\,|A) = \frac{N-n}{n}\sum_{g=1}^{G} N_g S^2_{y_{U_g}} + \frac{N(N-n)}{n^2(n-1)}\sum_{g=1}^{G}(N-N_h)S^2_{y_{U_g}}$$

■

Marco y Lucy

Una vez más volviendo a la población de empresas pertenecientes al sector industrial, suponga que se desea estimar el total de las características de interés mediante un estimador de post-estratificación que obedezca al modelo dado por la expresión (9.6.7), en donde las características de interés presentan diferentes comportamientos en la característica Nivel industrial que es constante (a través de cada subgrupo poblacional inducido por Nivel) y que supone un diferente comportamiento estructural a lo largo de toda la población, aunque homogéneo dentro de cada subgrupo. Suponga que se selecciona una muestra aleatoria simple de tamaño $n = 400$.

```
> data(Lucy)
> attach(Lucy)
> N <- dim(Lucy)[1]
> n <- 400
> sam <- sample(N,n)
> muestra <- Lucy[sam,]
> attach(muestra)
```

Para computar el estimador del total de las características de interés se define la matriz de información auxiliar, que en este caso particular corresponde a una matriz de variables «Dummy» cuyos vectores columna indican la pertenencia o no de una empresa seleccionada en la muestra al subgrupo poblacional. Por otro lado, se deben definir debidamente los parámetros de la función `GREG.SI` del paquete `TeachingSampling` según el modelo dado por (9.6.7)

```
> x <- Domains(Level)
> estima <- data.frame(Income, Employees, Taxes)
> pik <- rep(n/N,n)
> tx <- c(83,737,1576)
> b <- E.Beta(estima,x,pik,ck=1,b0=FALSE)
> GREG.SI(N,n,estima,x,tx, b, b0=FALSE)
                               Income    Employees        Taxes
Total estimado           1.030412e+06 1.502250e+05 2.824712e+04
Varianza                 1.933308e+08 8.282488e+06 6.452495e+05
coeficiente de variación 1.349396e+00 1.915747e+00 2.843738e+00
```

El siguiente código muestra que la propiedad de calibración se mantiene para este modelo. Por supuesto que la multiplicación de los valores observados de cada una de las características de interés con los pesos de calibración da como resultado la misma estimación que al utilizar la función `GREG.SI`.

```
> w <- Wk(x,tx,pik,ck=1,b0=FALSE)
> sum(x[,1]*w)
[1] 83
> sum(x[,2]*w)
[1] 737
> sum(x[,3]*w)
[1] 1576
> sum(1*w)
[1] 2396
```

A continuación se muestra la tabla de resultados que indica que asumiendo este modelo se gana mucha más precisión que en el caso en donde no se asume ningún modelo y se realizan las estimaciones directamente como en el caso del diseño de muestreo aleatorio simple. Nótese que para este caso particular, no se tuvo a la mano ninguna característica de información auxiliar sino que el modelo estuvo inducido por un simple vector de constantes que mejoro la estimación y la precisión.

Tabla 9.4: *Modelo de media post-estratificada: estimación de los totales de las características de interés.*

Variable	Total poblacional	Total estimado	cve %	Desv. %
Ingreso	1035217	1030412	1.34	-0.46
Empleados	151950	150225	1.91	-1.13
Impuestos	28654	28247	2.84	-1.42

9.6.4 Estimador de razón post-estratificada

Existe una gran cantidad de modelos que involucran subgrupos poblacionales. En este apartado vamos a considerar el estimador de razón post-estratificada que está sustentado en el modelo que lleva el mismo nombre, el cual supone la partición en G grupos de la población finita. De tal manera que $U = (U_1, U_2, \ldots, U_G)$. Se asume que es posible definir un modelo de razón en cada uno de los subgrupos U_g $g = 1, \ldots, G$. Así que se considera que la razón entre la característica de interés y la información auxiliar es constante dentro de cada subgrupo pero distinta entre cada subgrupo. Luego, $p = G$, $\mathbf{x}_k = \mathbf{d}_k x_k = (\underbrace{0, 0, \ldots, x_k, \ldots, 0, 0}_{G \texttt{ grupos}})'$ y $c_k = x_k$ para todo $k \in U_g$. De esta manera, el vector de totales de la característica de información auxiliar discreta está dado por

$$\mathbf{t_x} = (t_{x_{U_1}}, \ldots, t_{x_{U_g}}, \ldots, t_{x_{U_G}})' \tag{9.6.36}$$

donde

$$t_{x_{U_g}} = \sum_{U_g} x_k \qquad g = 1, \ldots, G. \tag{9.6.37}$$

y el vector de totales estimados de la característica de información auxiliar discreta está dado por

$$\hat{\mathbf{t}}_{\mathbf{x},\pi} = (\hat{t}_{x_{U_1},\pi}, \ldots, \hat{t}_{x_{U_g},\pi}, \ldots, \hat{t}_{x_{U_G},\pi})' \tag{9.6.38}$$

donde

$$\hat{t}_{x_{U_g},\pi} = \sum_{U_g} \frac{x_k}{\pi_k} \qquad g = 1, \ldots, G. \tag{9.6.39}$$

La formulación del modelo de razón post-estratificada está dada por

$$Y_k = \mathbf{X}_k' + \varepsilon_k \boldsymbol{\beta} \tag{9.6.40}$$

$$= \beta_g X_k + \varepsilon_k \qquad g = 1, \ldots, G. \tag{9.6.41}$$

Donde $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_g, \ldots, \beta_G)'$ y cada uno de los ε_k $k \in U_g$ son variables aleatorias independientes e idénticamente distribuidas con media cero y varianza σ_g^2 para $g = 1, \ldots, G$. Tenga en cuenta que $\mathbf{x}_k = (d_{1k}x_k, \ldots, d_{gk}x_k, \ldots, d_{Gk}x_k)'$ con

$$d_{gk} = \begin{cases} 1, & \text{si } k \in U_g \\ 0, & \text{en otro caso.} \end{cases} \tag{9.6.42}$$

Luego el modelo de super-población está dado por

$$\begin{aligned} E_\xi(Y_k) &= \mathbf{x}_k'\boldsymbol{\beta} = \beta_g x_k \\ Var_\xi(Y_k) &= x_k\sigma_g^2. \end{aligned} \tag{9.6.43}$$

La siguiente figura muestra el comportamiento de la relación entre la información auxiliar y la característica de interés bajo un modelo de razón post-estratificada.

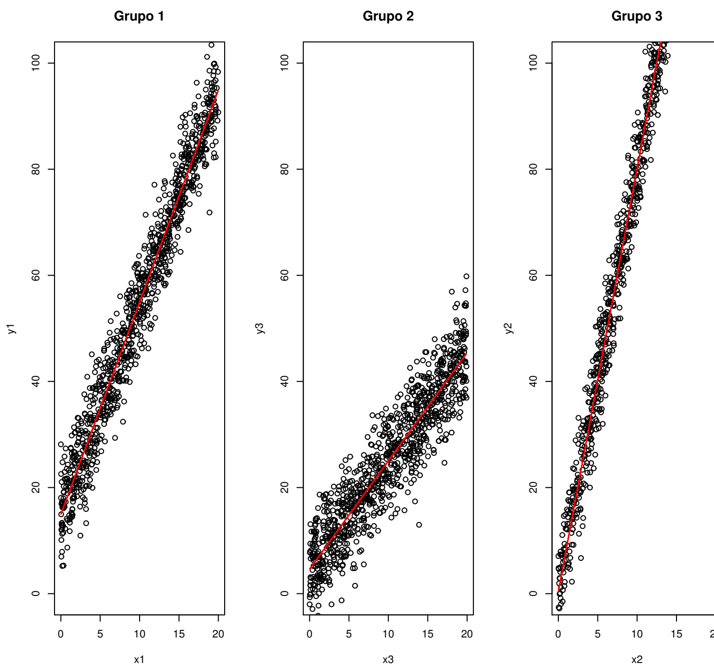


Figura 9.5: *Gráfico de dispersión de un modelo de razón post-estratificada.*

Särndal, Swensson & Wretman (1992) plantean dos ejemplos que sirven para describir cuándo utilizar este modelo de razón post-estratificada. El primero se refiere a la situación en donde la característica de interés y_k está dada por el *salario anual* que una empresa paga durante un año a todos sus empleados y en donde la característica de información auxiliar x_k es el *ingreso neto* que la empresa registra durante el mismo año. Bajo la anterior situación, es pertinente pensar que las empresas de un mismo grupo industrial tienen una razón y_k/x_k homogénea; pero las empresas de diferente grupo industrial no compartirán esa misma razón. El segundo ejemplo considera que la población, constituida por personas, puede ser particionada por subgrupos de edad o género. Si la característica de interés y_k es la cantidad de dinero anual que las personas gastan en recreación y si la característica de información auxiliar x_k es el salario que estas personas reciben

anualmente, entonces es razonable suponer que la razón y_k/x_k será la misma para personas del mismo subgrupo, pero distinta para personas de distinto subgrupo. Si se tuviese acceso a toda la población finita, el estimador del coeficiente de regresión $\boldsymbol{\beta}$ estaría dado por la minimización de la siguiente función de dispersión.

$$D = \sum_U \frac{(y_k - \mathbf{x}_k'\mathbf{B})^2}{x_k \sigma^2}. \tag{9.6.44}$$

Utilizando el resultado 8.4.2 y recurriendo a la ecuación (8.4.6), el estimador $\mathbf{B}$ en la población finita toma la siguiente forma

$$\mathbf{B} = (B_1, B_2, \ldots, B_G)' \tag{9.6.45}$$

En donde

$$B_g = \frac{\sum_{U_g} y_k}{\sum_{U_g} x_k} \qquad g = 1, \ldots, G \tag{9.6.46}$$

Por supuesto, como en la práctica sólo se tiene acceso a una muestra particular de la población finita, $\mathbf{B}$ debe ser estimado de tal manera que siguiendo el resultado 8.4.3. llegamos a la siguiente expresión

$$\hat{\mathbf{B}} = (\hat{B}_1, \hat{B}_2, \ldots, \hat{B}_G)' \tag{9.6.47}$$

En donde

$$\hat{B}_g = \left(\sum_{S_g} \frac{x_k}{\pi_k}\right)^{-1} \left(\sum_{S_g} \frac{y_k}{\pi_k}\right) = \frac{\sum_{S_g} \frac{y_k}{\pi_k}}{\sum_{S_g} \frac{x_k}{\pi_k}} = \frac{\hat{t}_{yU_g,\pi}}{\hat{t}_{yU_g,\pi}} \tag{9.6.48}$$

De esta manera, es posible ahora construir un estimador del total poblacional de la característica de interés. El siguiente resultado da cuenta de ello.

Resultado 9.6.6. *Bajo el modelo de razón post-estratificada, el estimador del total poblacional está dado por*

$$\hat{t}_{y,rpos} = \sum_{\substack{g=1 \\ n_h \neq 0}}^{G} N_g t_{x_{U_g}} \hat{B}_g \tag{9.6.49}$$

cuya varianza aproximada es

$$AVar(\hat{t}_{y,mpos}) = \sum\sum_U \Delta_{kl} \frac{E_k}{\pi_k} \frac{E_l}{\pi_l}. \tag{9.6.50}$$

con

$$E_k = y_k - B_g x_k \qquad g = 1, \ldots, G. \tag{9.6.51}$$

El estimador de la varianza es

$$\widehat{Var}(\hat{t}_{y,mpos}) = \sum\sum_S \frac{\Delta_{kl}}{\pi_{kl}} \frac{e_k}{\pi_k} \frac{e_l}{\pi_l} \tag{9.6.52}$$

con

$$e_k = y_k - \hat{B}_g x_k \qquad g = 1, \ldots, G. \tag{9.6.53}$$

Prueba. Como el estimador de razón post-estratificada es un caso particular del estimador general de regresión; entonces, por la configuración particular de $\mathbf{x}_k$ y adecuando la expresión (9.2.11), se tiene que

$$\begin{aligned}
\hat{t}_{y,rpos} &= \hat{t}_{y,\pi} + (\mathbf{t_x} - \hat{\mathbf{t}}_{\mathbf{x},\pi})'\hat{\mathbf{B}} \\
&= \hat{t}_{y,\pi} + (t_{x_{U_1}} - \hat{t}_{x_{U_1},\pi})\hat{B}_1 + \ldots + (t_{x_{U_G}} - \hat{t}_{x_{U_G},\pi})\hat{B}_G \\
&= \hat{t}_{y,\pi} - \sum_{g=1}^{G} \hat{t}_{x_{U_g},\pi}\hat{B}_g + \sum_{g=1}^{G} \hat{t}_{x_{U_g}}\hat{B}_g \\
&= \hat{t}_{y,\pi} - \sum_{g=1}^{G} \hat{t}_{y_{U_g},\pi} + \sum_{g=1}^{G} \hat{t}_{x_{U_g}}\hat{B}_g \\
&= \sum_{g=1}^{G} N_g \widetilde{y}_{S_g}
\end{aligned}$$

El cálculo de la varianza aproximada y la estimación de la varianza del estimador de razón son inmediatos al utilizar el resultado 9.2.3. Notando que

$$E_k = y_k - \mathbf{x}_k'\mathbf{B} = y_k - B_g x_k \qquad g = 1, \ldots, G.$$

Análogamente con e_k. ■

Por otro lado, fácilmente se puede demostrar que

$$\sum_s \frac{e_k}{\pi_k} = 0$$

puesto que

$$\begin{aligned}
\sum_S \frac{e_k}{\pi_k} &= \sum_S \frac{y_k}{\pi_k} - \sum_S \frac{\mathbf{x}_k}{\pi_k}'\hat{\mathbf{B}} \\
&= \hat{t}_{y,\pi} - \hat{t}_{x_{U_1},\pi}\frac{\hat{t}_{y_{U_1},\pi}}{\hat{t}_{x_{U_1},\pi}} - \cdots - \hat{t}_{x_{U_G},\pi}\frac{\hat{t}_{y_{U_G},\pi}}{\hat{t}_{x_{U_G},\pi}} \\
&= \hat{t}_{y,\pi} - \sum_{g=1}^{G} \hat{t}_{y_{U_g},\pi} = 0
\end{aligned}$$

Como consecuencia de lo anterior, es posible escribir al estimador de razón post-estratificada en la forma simplificada, teniendo en cuenta que $\hat{y}_k = \mathbf{x}_k'\hat{\mathbf{B}} = \hat{B}_g x_k$ cuando $k \in U_g$. Por tanto

$$\hat{t}_{y,mpos} = \sum_U \hat{y}_k = \sum_{g=1}^G \sum_{U_g} \hat{y}_k = \sum_{g=1}^G \hat{B}_g \sum_{U_g} x_k = \sum_{g=1}^G t_{x_{U_g}} \hat{B}_g$$

Además recurriendo a las expresiones (9.2.16) y (9.2.17) se tiene que

$$g_{ks} = 1 + \frac{t_{x_{U_g}} - \hat{t}_{x_{U_g},\pi}}{\hat{t}_{x_{U_g},\pi}} = \frac{t_{x_{U_g}}}{\hat{t}_{x_{U_g},\pi}} \qquad g = 1, \ldots, G. \tag{9.6.54}$$

Muestreo aleatorio simple

Resultado 9.6.7. *Bajo un diseño de muestreo aleatorio simple, el estimador de razón post-estratificada, su varianza aproximada y la estimación de la varianza están dados por*

$$\hat{t}_{y,rpos} = \sum_{g=1}^G t_{xg} \frac{\sum_{S_g} y_k}{\sum_{S_g} x_k} \tag{9.6.55}$$

$$Var_{MAS}(\hat{t}_{y,rpos}) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right) \sum_{g=1}^G W_g S^2_{E_{U_g}} \tag{9.6.56}$$

$$\widehat{Var}_{MAE}(\hat{t}_{y,rpos}) = \left(1 - \frac{n}{N}\right) \sum_{g=1}^G \left(\frac{\bar{x}_{U_g}}{\bar{x}_{S_g}}\right) \frac{N_g^2}{n_{s_g}} S^2_{e_{S_g}} \tag{9.6.57}$$

respectivamente. Donde $W_g = (N_g - 1)/(N - 1)$ *y* $S^2_{E_{U_g}}$ *es la varianza poblacional de la variable linealizada* E_k *en el post-estrato g-ésimo y* $S^2_{e_{S_g}}$ *es la varianza muestral de los valores* e_k *en el post-estrato g-ésimo. Nótese que*

$$E_k = y_k - \left(\frac{\sum_{U_g} y_k}{\sum_{U_g} y_k}\right) x_k \qquad \text{para todo } k \in U_g, \quad g = 1, \ldots, G$$

$$e_k = y_k - \left(\frac{\sum_{S_g} y_k}{\sum_{S_g} y_k}\right) x_k \qquad \text{para todo } k \in U_g, \quad G = 1, \ldots, G$$

Nótese que las expresiones correspondientes al resultado anterior son equivalentes a las del resultado 9.4.5 del estimador de razón separada bajo un diseño aleatorio estratificado.

Marco y Lucy

Retomando la población de empresas pertenecientes al sector industrial, suponga que se desea estimar el total de las características de interés Ingreso y Número de

Empleados mediante un estimador de post-estratificación que obedezca al modelo dado por la expresión (9.6.43), en donde las características de interés están relacionadas con una variable que es constante y que supone el mismo comportamiento estructural a lo largo de toda la población. Suponga que se selecciona una muestra aleatoria simple de tamaño $n = 400$.

```
> data(Lucy)
> attach(Lucy)
> N <- dim(Lucy)[1]
> n <- 400
> sam <- sample(N,n)
> muestra <- Lucy[sam,]
> attach(muestra)
```

Para computar el estimador del total de las características de interés se define la matriz de información auxiliar, que en este caso particular corresponde a la matriz de variables ≪Dummy≫ en donde cada columna está multiplicada por el vector de valores de la característica de información auxiliar Impuestos; es decir, una matriz de ceros, si la empresa no corresponde al subgrupo poblacional o valores en los reales si la empresa sí pertenece a tal subgrupo. Se utiliza la función `GREG.SI` del paquete `TeachingSampling` definiendo correctamente los parámetros según el modelo dado por (9.6.43)

```
> D <- Domains(Level)
> x <- D*Taxes
> estima <- data.frame(Income, Employees)
> pik <- rep(n/N,n)
> tx <- c(6251,16293,6110)
> b <- E.Beta(estima,x,pik,ck=1,b0=FALSE)
> GREG.SI(N,n,estima,x,tx, b, b0=FALSE)
                                Income    Employees
Total estimado            1.038163e+06 1.495162e+05
Varianza                  1.527426e+08 1.249836e+07
coeficiente de variación 1.190460e+00 2.364494e+00
```

El siguiente código muestra que la propiedad de calibración se mantiene para este modelo. Por supuesto que la multiplicación de los valores observados de cada una de las características de interés con los pesos de calibración da como resultado la misma estimación que al utilizar la función `GREG.SI`.

```
> w<-Wk(x,tx,pik,ck=1,b0=FALSE)
> sum(x[,1]*w)
[1] 6251
> sum(x[,2]*w)
[1] 16293
> sum(x[,3]*w)
[1] 6110
```

A continuación se muestra la tabla de resultados que indica que asumiendo este modelo se gana mucha más precisión que en el caso en donde e asume un modelo simple con particiones de subgrupos.

Tabla 9.5: *Modelo de razón post-estratificada: estimación de los totales de las características de interés.*

Variable	Total poblacional	Total estimado	cve %	Desv. %
Ingreso	1035217	1038163	1.19	0.28
Empleados	151950	152742	2.36	0.52

9.7 Múltiples modelos de regresión

En esta sección se trata de simular el ejemplo de la sección 7.9.1. de Särndal, Swensson & Wretman (1992), en donde se han construido cinco estimadores de regresión más el estimador de Horvitz-Thompson. Se realizó una simulación de Monte Carlo que se llevó a cabo con $Nsim = 5000$ repeticiones sobre la misma población (MU281). Se utilizaron dos variables auxiliares, $x_1 = CS82$ y $x_2 = SS82$, mientras que la variable e interés fue $y = \frac{RT85}{10000}$. Las estrategias de muestreo consisten en un diseño aleatorio simple sin reemplazo más el uso de los anteriores estimadores.

En cada corrida de la simulación se tomó una muestra aleatoria simple de tamaño $n = 100$ de $N = 281$ y se calcularon los siguientes estimadores:

$$t_\pi = N\bar{y}_S \tag{9.7.1}$$

$$t_{ra1} = \sum_s y_k \frac{\sum_U x_{1k}}{\sum_s x_{1k}} \tag{9.7.2}$$

$$t_{ra2} = \sum_s y_k \frac{\sum_U x_{2k}}{\sum_s x_{2k}} \tag{9.7.3}$$

$$t_{re2} = t_\pi + N\hat{b}_1(\bar{x}_{1s} - \bar{x}_{1U}) \tag{9.7.4}$$

$$t_{re2} = t_\pi + N\hat{b}_2(\bar{x}_{2s} - \bar{x}_{2U}) \tag{9.7.5}$$

$$t_r = t_\pi + N(\hat{B}_1(\bar{x}_{1s} - \bar{x}_{1U}) + \hat{B}_2(\bar{x}_{2s} - \bar{x}_{2U})) \tag{9.7.6}$$

donde

$$\hat{b}_j = \frac{\sum_s (x_{jk} - \bar{x}_{js})(y_k - \bar{y}_s)}{\sum_s (x_{jk} - \bar{x}_{js})^2} \quad j = 1, 2. \tag{9.7.7}$$

y

$$\hat{B} = (\hat{B}_0, \hat{B}_1, \hat{B}_2) = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'y \tag{9.7.8}$$

Se calcularon las diferentes varianzas descritas en la tabla 7.2. de Särndal, Swensson & Wretman (1992), para ello se tenían que computar los e_{ks} y los g_{ks} que para cada estimador, siguiendo el orden de las ecuaciones anteriores, estaban dados por

$$e_{k,\pi} = y_k - \bar{y}_S \tag{9.7.9}$$

$$e_{k,ra1} = y_k - \frac{\bar{y}}{\bar{x_1}} x_{1k} \tag{9.7.10}$$

$$e_{k,ra2} = y_k - \frac{\bar{y}}{\bar{x_2}} x_{2k} \tag{9.7.11}$$

$$e_{k,re1} = y_k - \hat{b}_1 x_{1k} \tag{9.7.12}$$

$$e_{k,re2} = y_k - \hat{b}_2 x_{2k} \tag{9.7.13}$$

$$e_{k,r} = y_k - \mathbf{x}_k \hat{B} \tag{9.7.14}$$

y los g_{ks} están dados por

$$g_{k,\pi} = NA \tag{9.7.15}$$

$$g_{k,ra1} = \frac{\bar{x_{1U}}}{\bar{x_{1s}}} \tag{9.7.16}$$

$$g_{k,ra2} = \frac{\bar{x_{2U}}}{\bar{x_{2s}}} \tag{9.7.17}$$

$$g_{k,re1} = 1 + n\left((\bar{x}_{1U} - \bar{x}_{1s})\frac{(x_{1k} - \bar{x}_{1s})}{\sum_s (x_{1k} - \bar{x}_{1s})^2}\right) \tag{9.7.18}$$

$$g_{k,re2} = 1 + n\left((\bar{x}_{2U} - \bar{x}_{2s})\frac{(x_{2k} - \bar{x}_{2s})}{\sum_s (x_{2k} - \bar{x}_{2s})^2}\right) \tag{9.7.19}$$

$$g_{k,r} = 1 + n(\bar{\mathbf{x}}_U - \bar{\mathbf{x}}_s)(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'_k \tag{9.7.20}$$

Para cada simulación se calculó el promedio de los intervalos de confianza en donde estaba t_y. La simulación fue programada en el paquete estadístico R. Los resultados de la simulación se presentan a continuación.

Tabla 9.6: *Resultados de la simulación de Monte Carlo para el ejercicio propuesto en Särndal, Swensson & Wretman (1992).*

Est	$\hat{t}$	$S^2_{\hat{t}}$	$\hat{V}_g$	ECR_g	$\hat{V}_{sim}$	ECR_{sim}	AV
$\hat{t}_\pi$	5.307	0.056	0.051	92.68	0.050	91.80	0.051
$\hat{t}_{ra1}$	5.313	0.122	0.120	93.46	0.120	93.56	0.121
$\hat{t}_{ra2}$	5.315	0.142	0.140	93.98	0.141	94.06	0.141
$\hat{t}_{re1}$	5.309	0.120	0.141	94.88	0.140	94.78	0.136
$\hat{t}_{re2}$	5.303	0.123	0.361	99.16	0.359	99.06	0.349
$\hat{t}_r$	5.307	0.056	0.051	92.68	0.050	91.80	0.051

Nótese que la tabla 9.6. da los resultados de la simulación. Ésta refleja en la primera columna el nombre del estimador; en la segunda y en la tercera columna se da el promedio y la varianza para las 5000 repeticiones; la cuarta y quinta columna indican el promedio de las estimaciones de la varianza utilizando la expresión (9.2.39) y la tasas de cobertura en todas las 5000 simulaciones para los intervalos de confianza inducidos por esta varianza; por último, las columnas sexta y séptima representan el promedio de las estimaciones de la varianza utilizando la expresión (9.2.30) y la tasas de cobertura en cada una de las 5000 simulaciones para los intervalos de confianza inducidos por esta varianza.

A continuación se presenta el programa de simulación que entrega los resultados para la última fila de la tabla. Para obtenerlo, se necesito crear dos funciones que se explican a continuación. Nótese que para obtener los resultados de las otras filas, se be modificar el código de acuerdo a las especificaciones particulares de cada caso.

```
> library(sampling)
> data(MU284)
> MU281 <- MU284[MU284$RMT85 <= 3000,]
> attach(MU281)

> #Fija una semilla para obtener los mismos resultados
> set.seed(7654321)
> #Tamaño de la simulación de Monte Carlo
> Nsim<-5000

> #Función que calcula las estimaciones
> r <- function(Y,X1,X2,N,n){
+ ty <- sum(Y)
+ sam <- sample(N,n)
+ y <- Y[sam]
+ x1 <- X1[sam]
+ x2 <- X2[sam]
+ X <- cbind(1,X1,X2)
+ x <- cbind(1,x1,x2)
+ beta <-solve(t(x)%*%x, t(x)%*%y)
+ beta1 <- beta[2]
+ beta2 <- beta[3]
+ Beta <- solve(t(X)%*%X, t(X)%*%Y)
+ #Calcula el estimador de Horvitz-Thompson
+ tHT <- N*mean(y)
+ #Calcula el estimador de regresión múltiple
+ tr <- tHT+N*((beta1)*(mean(X1)-mean(x1))+(beta2)*(mean(X2)-mean(x2)))
+ er <- y - x%*%beta
+ Er <- Y - X%*%Beta
+ #Calcula de los pesos gk
+ gr <- 1+n*(colMeans(X)-colMeans(x))%*%solve(t(x)%*%x)%*%t(x)
+ gr <- as.vector(gr)
+ #Calcula la varianza del estimador con los gk
+ Vgr <- k*sum(gr^2*er^2)
+ Crgr=0
+ LBgr <- tr-1.96*sqrt(Vgr)
+ UBgr <- tr+1.96*sqrt(Vgr)
+ #Calcula la cobertura del estimador de regresión múltiple
+ if(ty>=LBgr & ty<=UBgr) Crgr=100
+ #Calcula la varianza del estimador sincon los gk
+ Vsr <- k*sum(er^2)
+ Crsr=0
+ LBsr <- tr-1.96*sqrt(Vsr);UBsr <- tr+1.96*sqrt(Vsr)
+ #Calcula la cobertura del estimador de regresión múltiple
+ if(ty>=LBsr & ty<=UBsr) Crsr=100
+ AVr <- K*sum(Er^2)
+ todo<-c(tr, Vgr, Crgr, Vsr, Crsr, AVr)
+ return(todo)
+ }

> Y <- RMT85/10000
> X1 <- CS82
> X2 <- SS82
> N <- 281
> n <- 100
> k <- (N^2)*(1-(n/N))/(n^2-n)
> K <- (N^2)*(1-(n/N))/(n*(N-1))

> RES <- matrix(0,Nsim,6)
> for(j in 1:Nsim){
+ RES[j,] <- r(Y,X1,X2,N,n)
+ }
```

```
> tbar <- mean(RES[,1])
> tvar <- var(RES[,1])
> vg <- mean(RES[,2])
> ECRg <- mean(RES[,3])
> vsim <- mean(RES[,4])
> ECRs <- mean(RES[,5])
> AV <- mean(RES[,6])

> c(tbar,tvar,vg,ECRg,vsim,ECRs,AV)
[1] 5.307 0.056 0.051 92.680 0.050 91.800 0.051
```

9.8 Ejercicios

9.1 Realice el ejercicio lexicográfico del Ejemplo 9.2.1.

9.2 Realice la demostración del resultado 9.4.4.

9.3 Demuestre que la estimación de la varianza del estimador de razón bajo muestreo aleatorio simple se puede expresa como:

$$\widehat{Var}_{MAS}(\hat{t}_{y,ra}) = \frac{N^2}{n}\left(1-\frac{n}{N}\right)\frac{\sum_S(y_k-\hat{B}x_k)^2}{n-1}$$

con $\hat{B} = \sum_s y_k / \sum_s y_k$. Luego, demuestre que esta expresión puede ser escrita como

$$\widehat{Var}_{MAS}(\hat{t}_{y,ra}) = \frac{N^2}{n}\left(1-\frac{n}{N}\right)(S^2_{y_s} + \hat{B}^2 S^2_{x_s} - 2\hat{B}S^2_{xy_s})$$

9.4 Si se utiliza un diseño de muestreo aleatorio simple con un estimador de post-estratificación para el total poblacional,

(a) Demuestre que $\bar{E}_U = \bar{e}_s = 0$.

(b) Proponga una expresión para estimar la varianza de los estimadores del total en cada post-estrato.

9.5 Mediante fuentes externas, se sabe que en la población del ejercicio 8.9 hay 24796 automoviles y 4968 efectivos militares. Cada municipio se categoriza (**CAT**) en urbano (**CAT**=1) o rural (**CAT**=0) y existen 22 municipios pertenecientes a la categoría urbana y los restantes 27 pertenecen a la categoría rural. En los municipios urbanos existen en total 11132 vehículos y en los municipios rurales hay 13664 vehículos.

(a) Estime el estimador del total de habitantes usando un modelo de media común de **HAB** contra **VEH**. Estime la varianza y calcule el coeficiente de variación.

(b) Estime el estimador del total de habitantes usando un modelo de razón de **HAB** contra **VEH**. Estime la varianza y calcule el coeficiente de variación.

(c) Estime el estimador del total de habitantes usando un modelo de regresión simple con intercepto de **HAB** contra **MIL**. Estime la varianza y calcule el coeficiente de variación.

(d) Estime el estimador del total de habitantes usando un modelo de regresión múltiple con intercepto de **HAB** contra **MIL** y **VEH**. Suponga que el modelo es homoscedástico. Estime la varianza y calcule el coeficiente de variación.

(e) Estime el estimador del total de habitantes usando un modelo de media post-estratificada de **HAB** contra **CAT**. Estime la varianza y calcule el coeficiente de variación.

(f) Estime el estimador del total de habitantes usando un modelo de razón post-estratificada de **HAB** contra **MIL** para post-estratos construidos mediante **CAT**. Estime la varianza y calcule el coeficiente de variación.

9.6 Considere una población de 2010 predios agropecuarios. Asuma que el proceso de estratificación fue llevado a cabo de la siguiente manera: el primer estrato está conformado por todos aquellos predios que tienen un área cultivada menor de 160 hectáreas (Estrato 1); el otro estrato está conformado por aquellos predios con un área cultivada mayor de 160 hectáreas (Estrato 2). Para esta configuración, en el primer estrato se seleccionó una muestra aleatoria simple de $n_1 = 70$ y en el segundo estrato se seleccionó una muestra aleatoria simple de $n_1 = 30$. Los datos obtenidos se presentan en la siguiente tabla.

Estrato	N_h	n_h	$\bar{x}_U$	$\bar{y}_s$	$\bar{x}_s$	$S^2_{y_s}$	$S^2_{e_s}$
1	1580	70	84	19.4	82.5	312	20.5
2	430	30	241	51.6	244.85	922	73.6
Total	2010	100	–	–	–	620	76.1

Compare las varianzas estimadas y coeficientes de variación estimados para el estimador del total poblacional en las siguientes estrategias de muestreo:

(a) Muestreo aleatorio simple con el estimador de Horvitz-Thompson

(b) Muestreo aleatorio simple con el estimador de razón

(c) Muestreo aleatorio estratificado con el estimador de Horvitz-Thompson

(d) Muestreo aleatorio estratificado con el estimador de razón separada

(e) Muestreo aleatorio estratificado con el estimador de razón combinada

9.7 Sustente o refute las siguientes afirmaciones

(a) En la determinación de un estimador de regresión general de regresión, si en lugar de los verdaderos totales de las características de información auxiliar, se dispone de estimaciones gruesas dadas por fuentes administrativas, el estimador general de regresión será de varianza muy alta.

(b) En la medida que el tamaño de muestra aumenta, el sesgo del estimador general de regresión decrece.

(c) En la medida que el tamaño de muestra aumenta, la varianza del estimador general de regresión decrece.

(d) Al momento de decidir entre un diseño estratificado o un estimador de post-estratificación, el carácter multipropósito de los estudios induce a preferir el diseño estratificado.

(e) Si el intercepto es importante en la descripción de la nube de puntos, el utilizar un estimador de regresión que no contemple la presencia del intercepto incrementa la varianza y el sesgo de la estrategia.

(f) Al planear un diseño estratificado y utilizar el estimador general de regresión, se debe buscar que los estratos se construyan con base en la similitud de valores de los residuales del ajuste de regresión.

(g) La magnitud de la varianza del estimador general de regresión siempre depende de la bondad del ajuste entre la característica de interés y las características de información auxiliar.

Capítulo 10

Estimadores de calibración

> La calibración [como proceso] se ha establecido como un importante instrumento metodológico en la producción de grandes masas de estadísticas. La mayoría de agencias estadísticas han desarrollado software especialmente diseñado para calcular las ponderaciones resultantes, usualmente calibradas a la información auxiliar disponible en registros administrativos y otras fuentes precisas.
>
> Carl-Erik Särndal (2008)

El proceso de calibración es el tema principal de los más recientes artículos publicados acerca de estimación en poblaciones finitas y muestreo. Este fenómeno se presenta debido a que la calibración provee una forma sistemática para la incorporación de la información auxiliar en la etapa de estimación en una encuesta. Un estimador de calibración es aquel estimador lineal que tiene la agradable propiedad de la representatividad bajo cualquier diseño de muestreo; aunque el término calibración es nuevo, hay autores que coinciden en afirmar que han usado calibración desde mucho tiempo atrás, antes de conocer este proceso con éste nombre.

Como Särndal (2007) afirma, el ítem más importante en la calibración, como proceso sistemático de estimación, es la existencia de información auxiliar. Si no hay información auxiliar no hay nada a lo que se pueda calibrar, y por tanto no habrán estimadores de calibración que aplicar. Como se verá a lo largo del capítulo, los estimadores generales de regresión pueden arrojar los mismos resultados que los estimadores de calibración; sin embargo, el espíritu y la esencia de su aplicación tienen direcciones marcadamente diferentes.

¿Pero qué es un estimador de calibración? ¿cuál es su esencia?. A continuación una breve descripción de este método:

1. Suponga que se tiene acceso a un vector de información auxiliar, $\mathbf{x}_k = (x_{1k}, x_{2k}, \ldots, x_{pk})$, de p variables auxiliares y conocido para los individuos seleccionados en la muestra.

2. Además, por registros administrativos u otras fuentes de confianza, se tiene el conocimiento del total del vector de información auxiliar $\mathbf{t_X} = \sum_{k\in U} \mathbf{x}_k$.

3. El propósito del estudio es estimar el total de la característica de interés usando la información dada por $\mathbf{x}_k \quad k \in S$.
4. Aunque el estimador de Horvitz-Thompson es insesgado, se requiere que las estimaciones cumplan con la siguiente restricción dada por

$$\sum_{k \in S} w_k \mathbf{x}_k = \mathbf{t_X}$$

y conocida como ecuación de calibración.

5. La idea consiste en buscar estos pesos w_k tan cercanos como sea posible al inverso de la probabilidad de inclusión del k-ésimo elemento $d_k = 1/\pi_k$

Aunque el concepto de calibración es nuevo en la teoría de muestreo, la esencia del método y el espíritu práctico de obtener estimaciones que ajusten exactamente con totales conocidos no es nuevo. De hecho, este método se ha utilizado, y algunos investigadores lo están utilizando, sin saber que se llama calibración. Este fue el caso de Deming & Stephan (1940) quienes abordaron este tema utilizando tablas de contingencia con estimaciones internas y totales marginales conocidos. Ellos fueron los pioneros del **proceso iterativo de ajuste proporcional** o IPFP, por sus siglas en inglés.

10.1 IPFP

Suponga que existen dos variables cualitativas que dividen la poblaciones en subgrupos poblacionales. Por un lado una variable permite dividir la población en H subgrupos poblacionales, $U_{1\cdot}, \ldots, U_{h\cdot}, \ldots, U_{H\cdot}$, y por otro lado una variable que permite dividir la población en G subgrupos poblacionales, $U_{\cdot 1}, \ldots, U_{\cdot g}, \ldots, U_{\cdot G}$. Como resultado la población se particiona en $H \times G$ subgrupos poblacionales como lo muestra la siguiente tabla.

Tabla 10.1: *Distribución de la población en la tabla de contingencia.*

U_{11}	$\cdots$	U_{1g}	$\cdots$	U_{1G}	$U_{1\cdot}$
$\vdots$		$\vdots$		$\vdots$	$\vdots$
U_{h1}	$\cdots$	U_{hg}	$\cdots$	U_{hG}	$U_{h\cdot}$
$\vdots$		$\vdots$		$\vdots$	$\vdots$
U_{H1}	$\cdots$	U_{Hg}	$\cdots$	U_{HG}	$U_{H\cdot}$
$U_{\cdot 1}$	$\cdots$	$U_{\cdot g}$	$\cdots$	$U_{\cdot G}$	U

Los tamaños de los subgrupos poblacionales se definen así: $N_{hg} = \#U_{hg}$, $N_{h\cdot} = \#U_{h\cdot}$, $N_{\cdot g} = \#U_{\cdot g}$. Nótese que se tiene que

$$N = \sum_{h=1}^{H} N_{h\cdot} = \sum_{g=1}^{G} N_{\cdot g}. \tag{10.1.1}$$

Además de esto, los totales de las celdas de la tabla de contingencia siguen la siguiente relación:

Tabla 10.2: *Distribución de los tamaños poblacionales en la tabla de contingencia.*

N_{11}	$\cdots$	N_{1g}	$\cdots$	N_{1G}	$N_{1\cdot}$
$\vdots$		$\vdots$		$\vdots$	$\vdots$
N_{h1}	$\cdots$	N_{hg}	$\cdots$	N_{hG}	$N_{h\cdot}$
$\vdots$		$\vdots$		$\vdots$	$\vdots$
N_{H1}	$\cdots$	N_{Hg}	$\cdots$	N_{HG}	$N_{H\cdot}$
$N_{\cdot 1}$	$\cdots$	$N_{\cdot g}$	$\cdots$	$N_{\cdot G}$	N

Después de la recolección y observación de los datos en la encuesta, se tiene la estimación definitiva de los totales de cada una de las celdas internas y de las celdas marginales. Así, $\hat{N}_{hg}$ corresponde a la estimación de N_{hg}, $\hat{N}_{h\cdot}$ corresponde a la estimación de $N_{h\cdot}$, $N_{\cdot g}$ corresponde a la estimación de $N_{\cdot g}$ y por último, $\hat{N}$ corresponde a la estimación de N. De está manera, es posible utilizar el estimador de Horvitz-Thompson, definiendo

$$N_{\cdot g} = \sum_{k \in U} z_{hk} \qquad N_{h\cdot} = \sum_{k \in U} z_{gk}.$$

Donde,

$$z_{hk} = \begin{cases} 1 & \text{si } k \in U_{h\cdot} \\ 0 & \text{en otro caso} \end{cases} \qquad z_{gk} = \begin{cases} 1 & \text{si } k \in U_{\cdot g} \\ 0 & \text{en otro caso} \end{cases}$$

Al utilizar el estimador de Horvitz-Thompson se garantiza el insesgamiento y se tiene la relación dada por la siguiente tabla

Tabla 10.3: *Distribución de los tamaños poblacionales estimados en la tabla de contingencia.*

$\hat{N}_{11}$	$\cdots$	$\hat{N}_{1g}$	$\cdots$	$\hat{N}_{1G}$	$\hat{N}_{1\cdot}$
$\vdots$		$\vdots$		$\vdots$	$\vdots$
$\hat{N}_{h1}$	$\cdots$	$\hat{N}_{hg}$	$\cdots$	$\hat{N}_{hG}$	$\hat{N}_{h\cdot}$
$\vdots$		$\vdots$		$\vdots$	$\vdots$
$\hat{N}_{H1}$	$\cdots$	$\hat{N}_{Hg}$	$\cdots$	$\hat{N}_{HG}$	$\hat{N}_{H\cdot}$
$\hat{N}_{\cdot 1}$	$\cdots$	$\hat{N}_{\cdot g}$	$\cdots$	$\hat{N}_{\cdot G}$	$\hat{N}$

Hasta el momento, se ha cumplido con el objetivo de estimar las celdas internas y las marginales de la tabla de contingencia. Sin embargo, suponga que, debido a registros administrativos u otras fuentes de confianza, es posible tener acceso a los totales de las celdas marginales tanto por columnas como por filas. Es decir, suponga que $N_{\cdot g}$, $g = 1, \ldots, G$ y $N_{h\cdot}$, $g = 1, \ldots, G$ son conocidos.

Bajo el anterior supuesto, es posible construir un algoritmo que ajuste las estimaciones de las celdas internas y que tenga la agradable propiedad que, finalizado

el algoritmo, al sumar por filas y columnas, las estimaciones correspondan a los totales conocidos de las celdas marginales. Este método de estimación basado en un algoritmo muy simple se conoce como **proceso iterativo de ajuste proporcional** o IPFP, por sus siglas en inglés, y fue propuesto por Deming & Stephan (1940).

10.1.1 Algoritmo

Aunque simple e intuitivo, el siguiente algoritmo es muy potente y tiene la buena propiedad de converger muy rápidamente si la tabla de contingencia no tiene valores nulos en sus celdas internas y si los totales marginales conocidos tienen sentido con la puesta en marcha de la encuesta.

1. Inicializar con

$$N_{hg}^{(0)} = \hat{N}_{hg} \qquad g = 1, \ldots, G, h = 1, \ldots, H$$

2. Para $t = 1, 2, 3, \ldots$

$$N_{hg}^{(2t-1)} = N_{hg}^{(2t-2)} \frac{N_{h\cdot}}{\sum_{g=1}^{G} N_{hg}^{(2t-2)}} \qquad g = 1, \ldots, G, h = 1, \ldots, H$$

$$N_{hg}^{(2t)} = N_{hg}^{(2t-1)} \frac{N_{\cdot g}}{\sum_{h=1}^{H} N_{hg}^{(2t-1)}} \qquad g = 1, \ldots, G, h = 1, \ldots, H$$

A simple vista, un defecto significativo de este método es que no tiene en cuenta el diseño de muestreo del cual provienen los datos para calibrar con respecto a la información auxiliar conocida. Sin embargo, como se verá en las próximas secciones, Deville & Särndal (1992) y Deville, Särndal & Sautory (1993) probaron que efectivamente, el proceso iterativo de ajuste proporcional se podía tratar como un caso especial de los estimadores de calibración bajo el espíritu del numeral 5 de la introducción. A los estimadores de calibración que surgen bajo este marco de referencia se les conoce con el nombre de estimadores generalizados de raking.

10.1.2 Marco y Lucy

Volviendo con nuestra población de empresas del sector industrial, se sabe que las variables cualitativas Nivel y SPAM conforman una partición de la población. Por un lado, la variable Nivel, divide a la población en tres subgrupos de acuerdo a características de la empresa, a saber: Grande, Mediana y Pequeña. Por otro lado, la variable SPAM, divide a la población en dos subgrupos poblacionales, de acuerdo a sus estrategias publicitarias, así: `SPAM.SI` y `SPAM.NO`. En total la población se divide en $2 \times 3 = 6$ subgrupos poblacionales.

Ahora, suponga que se ha planeado un diseño de muestreo aleatorio simple con un tamaño de muestra $n = 400$ y que se desea estimar el total de empresas por grupo industrial, el total de empresas que usan y no usan SPAM y su respectiva anidación interna en la tabla de contingencias, como lo muestra la siguiente tabla.

Tabla 10.4: *Tabla de contingencia para SPAM.*

	SPAM.NO	SPAM.SI	Total
Grande	N_{11}	N_{12}	$N_{1\cdot}$
Mediana	N_{21}	N_{22}	$N_{2\cdot}$
Pequeña	N_{31}	N_{32}	$N_{3\cdot}$
Total	$N_{\cdot 1}$	$N_{\cdot 2}$	N

En primer lugar, con ayuda de la función `S.SI` perteneciente al paquete `TeachingSampling`, se debe seleccionar una muestra probabilística de tamaño $n = 400$.

```
> data(Lucy)
> attach(Lucy)
> N <- dim(Lucy)[1]
> n <- 400
> sam <- S.SI(N,n)
> muestra <- Lucy[sam,]; attach(muestra)
```

Una vez que se ha observado y recolectado la información de cada una de las empresas seleccionadas en la muestra, se utiliza la función `Domains` del paquete `TeachingSampling` para obtener dos matrices, `SPAM.no` y `SPAM.si`, que indican la pertenencia o no de cada empresa seleccionada en la muestra a cada uno de los tres niveles del sector industrial.

```
> estima   <- data.frame(Domains(Level))
> Dominios <- data.frame(Domains(SPAM))
> SPAM.no <- Dominios[,1]*estima
> SPAM.si <- Dominios[,2]*estima
```

A continuación se muestran los cinco primeros elementos de las dos matrices creadas.

```
> SPAM.no                    |    > SPAM.si
    Big Medium Small         |        Big Medium Small
1     0      0     1         |    1     0      0     0
2     0      0     0         |    2     0      0     1
3     0      0     0         |    3     0      0     1
4     0      0     1         |    4     0      0     0
5     0      0     0         |    5     0      0     1
...                          |     ...
```

Para estimar los totales marginales correspondientes a las variables `Level` y `SPAM`, utilizamos la función `E.SI` del paquete `TeachingSampling`, la cual se aplica sobre los objetos `estima` y `dominios`, creados en el paso anterior.

```
> E.SI(N,n,estima)
                               Big      Medium        Small
```

```
Total estimado          107.82000  676.870000 1611.310000
Varianza                515.09857 2429.488304 2639.842691
coeficiente de variación  21.04970    7.282025    3.188673

> E.SI(N,n,Dominios)
                                no          yes
Total estimado           868.550000  1527.450000
Varianza                2769.890846  2769.890846
coeficiente de variación    6.059496     3.445596
```

Para estimar las celdas internas de la tabla de contingencia, utilizamos la función `E.SI` del paquete `TeachingSampling`, la cual se aplica sobre las matrices `SPAM.no` y `SPAM.si`, creadas anteriormente.

```
> E.SI(N,n,SPAM.no)
                              Big      Medium       Small
Total estimado           35.94000   245.59000  587.020000
Varianza                177.09322 1102.63755 2217.111277
coeficiente de variación  37.02737    13.52090    8.021228

> E.SI(N,n,SPAM.si)
                              Big       Medium      Small
Total estimado           71.88000  431.280000 1024.29000
Varianza                348.79275 1769.134340 2933.49981
coeficiente de variación  25.98218     9.752615    5.28774
```

Tabla 10.5: *Estimación de Horvitz-Thompson para la tabla de contingencia de SPAM.*

	SPAM.NO	SPAM.SI	Total
Grande	35.94	71.88	107.82
Mediana	245.59	431.28	676.87
Pequeña	587.02	1024.29	1611.31
Total	868.55	1527.45	2396

Por tanto, la estimación de Horvitz-Thompson bajo muestreo aleatorio simple está dada por la tabla 10.5. Ahora, suponga que, debido a registros administrativos u otras fuentes de confianza, es posible conocer el valor de los totales marginales para `Level` y `SPAM`; dadas por 83 empresas grandes, 737 empresas medianas y 1576 empresas pequeñas, para la variable `Level` y por 937 empresas que no utilizan SPAM y 1459 empresas que sí utilizan SPAM, para la variable `SPAM`. Es posible, entonces, utilizar el procedimiento iterativo de ajuste proporcional para calibrar las estimaciones internas de la tabla de contingencia para que ajusten exactamente a los valores poblacionales conocidos. Lo primero que se debe hacer, se debe crear la tabla de contingencia en R.

```
> Tab <- matrix(c(35.94,245.59,587.02,71.88,431.28,1024.29),3,2)
> rownames(Tab) <- c("Grande", "Mediana","Pequena")
> colnames(Tab) <- c("SPAM.NO", "SPAM.SI")
```

Una vez creada la tabla de contingencias, procedemos a implementar el algoritmo mediante la función `IPFP` del paquete `TeachingSampling`. Esta función consta de cuatro argumentos. El primer argumento es `Tab`, concerniente a la tabla de contingencias resultante de la estimación mediante el diseño probabilístico. El segundo argumento es `Col` y es un vector que contiene los totales marginales (poblacionales y conocidos) de las columnas de la tabla de contingencia. El tercer argumento es `Row` y es un vector que contiene los totales marginales (poblacionales y conocidos) de las filas de la tabla de contingencia. Por último `tol`, que por defecto es equivalente a 0.00001, corresponde a la tolerancia del algoritmo. La función `IPFP` arroja como resultado una tabla de contingencias calibrada según los argumentos `Col` y `Tol`. Para este ejemplo particular, se tiene la siguiente salida:

```
> Col <- c(937,1459)
> Row <- c(83,737,1576) > IPFP(Tab,Col,Row,tol=0.00001)

            SPAM.NO       SPAM.SI  Row.est
Grande     29.93189      53.06811       83
Mediana   288.25691     448.74309      737
Pequena   618.81120     957.18880     1576
Col.est   937.00000    1459.00000     2396
```

A continuación se encuentran las tablas comparativas de las estimaciones calibradas mediante el proceso iterativo de ajuste proporcional y la información correspondiente a los totales poblacionales, respectivamente.

Tabla 10.6: *Distribución poblacional (no conocida) para la tabla de contingencia de SPAM.*

	SPAM.NO	SPAM.SI	Total
Grande	29.93	53.06	83
Mediana	288.25	448.74	737
Pequeña	618.81	957.18	1576
Total	937	1459	2396

Tabla 10.7: *Estimación IPFP de calibración para la tabla de contingencia de SPAM.*

	SPAM.NO	SPAM.SI	Total
Grande	26	57	83
Mediana	291	446	737
Pequeña	620	956	1576
Total	937	1459	2396

Nótese que la diferencia relativa es muy pequeña y que las estimaciones se acercan a la verdad. En estos términos relativos, esta estimación resulta mejor que la inducida por el estimador de Horvitz-Thompson.

10.2 Fundamentos teóricos

Como se estableció en la anterior sección, los estadísticos han intentado utilizar la incorporación de información auxiliar para mejorar las estimaciones de la encuesta. Es así como el estimador de regresión en todas sus posibles formas, requiere el conocimiento del total de un vector de variables auxiliares. Como Deville & Särndal (1992) lo explican, los estimadores de calibración son una familia o clase de estimadores que tienen una forma muy atractiva y que se caracteriza por usar pesos calibrados, los cuales son tan cercanos como sea posible a los pesos originales o inversos de la probabilidad de inclusión del elemento seleccionado en la muestra y además estos estimadores de calibración respetan un conjunto de restricciones, las ecuaciones de calibración.

Considere una población finita $U = \{1, \ldots, k, \ldots, N\}$, de la cual se ha seleccionado una muestra probabilística s ($s \subseteq U$) inducida por un diseño de muestreo $p(\cdot)$. Luego, $p(s)$ es la probabilidad de que la muestra s haya sido seleccionada. Se asume que las probabilidades de inclusión de primer y segundo orden son estrictamente positivas.

Sea y_k el valor de la característica de interés para el k-ésimo individuo de la población, el cual también tiene asociado un vector de valores auxiliares dado por $\mathbf{x}_k = (x_{1k}, x_{2k}, \ldots, x_{pk})$. Nótese que y_k y $\mathbf{x}_k$ se observan y se conocen para todos los elementos en la muestra. Además, se asume que se conoce, mediante registros administrativos u otras fuentes de confianza, el total poblacional del vector de información auxiliar $\mathbf{t_x} = \sum_{k \in U} \mathbf{x}_k$.

Como en la mayoría de situaciones que se presentan en este libro, el objetivo es estimar el total poblacional de la característica de interés, t_y. Sin embargo, el estimador de t_y debe ser un estimador lineal de la forma

$$\hat{t}_S(y) = \sum_{k \in S} w_k y_k, \tag{10.2.1}$$

Nótese que el estimador de Horvitz-Thompson toma la anterior forma pues

$$\hat{t}_{y,\pi} = \sum_{k \in S} \frac{y_k}{\pi_k} = \sum_{k \in S} d_k y_k, \tag{10.2.2}$$

Además de la linealidad, la familia de estimadores de calibración debe inducir una estrategia de muestreo representativa para cualquier diseño de muestreo $p(\cdot)$. Es decir, se deben construir unos nuevos pesos w_k, que sean tan cercanos como sea posible a $d_k = 1/\pi_k$ considerando alguna métrica y, que además cumplan con las ecuaciones de calibración

$$\sum_{k \in S} w_k \mathbf{x}_k = \mathbf{t_X}. \tag{10.2.3}$$

Nótese que los pesos w_k dependen de S. Por ejemplo bajo el diseño MAS, el estimador de razón se puede escribir como

$$\hat{t}_{yr} = \hat{t}_{y\pi}\frac{t_x}{\hat{t}_{x\pi}} = \sum_{k\in S}\frac{N}{n}\frac{\bar{x}_U}{\bar{x}_S}y_k = \sum_{k\in S}w_k y_k$$

Además los pesos w_k tienen la propiedad de calibración puesto que

$$\sum_S w_k x_k = \sum_S \frac{N}{n}\frac{\bar{x}_U}{\bar{x}_S}x_k = N\frac{\bar{x}_U}{\bar{x}_S}\sum_s \frac{x_k}{n} = N\bar{x}_U = t_x$$

Dado que existe una variedad de estimadores que cumplen la restricción (10.2.3), se deben encontrar unos pesos w_k que tengan las siguientes propiedades (Estevao, Särndal & Sautory 2000)

1. **Consistencia**: un sistema de pesos o ponderaciones que satisfaga (10.2.3) es atractivo, porque reproduce exactamente el total poblacional conocido para cada variable auxiliar.
2. **Cercanía a los pesos básicos**: los pesos básicos $d_k = 1/\pi_k$ tienen la atractiva propiedad de inducir estimaciones insesgadas con respecto al diseño de muestreo utilizado. Se quiere que cualquier desviación de estos pesos sea pequeña para preservar esta propiedad, al menos aproximadamente o asintóticamente.
3. **Control sobre los totales de las variables auxiliares**: lo que dice la intuición es que entre más variables auxiliares sean usadas en el proceso de calibración, entonces mejor la estimación. Este argumento intuitivo es soportado por la teoría; de esta manera, Estevao, Särndal & Sautory (2000, sec. 6.) demuestran que la varianza de un estimador de calibración decrece mientras más variables auxiliares sean tenidas en cuenta en la calibración.

10.3 Construcción

Para construir estos nuevos pesos w_k, se debe minimizar una pseudo-distancia[1] $G(w_k/d_k)$ entre w_k y d_k en toda la muestra. Éste se puede tomar como un problema de optimización de la distancia en toda la muestra dada por

$$\sum_{k\in S} d_k\frac{G(w_k/d_k)}{q_k} \tag{10.3.1}$$

sujeto a la restricción (10.2.3). Donde, q_k $(k \in S)$ forman un conjunto de ponderaciones conocidas y estrictamente positivos. Acerca de la pseudo-distancia $G(w_k/d_k)$, se supone que

[1]Una función de distancia $D(x_1, x_2)$ debe cumplir con las siguientes propiedades: i) ser estrictamente positiva (no negativa), decir que $D(x_1, x_2) \geq 0$; ii) $D(x_1, x_2) = 0$ únicamente cuando $x_1 = x_2$; iii) ser simétrica, es decir $D(x_1, x_2) = D(x_2, x_1)$; cumplir con la desigualdad triángular, es decir $D(x_1, x_3) \leq D(x_1, x_2) + D(x_2, x_3)$. La función $G(w_k/d_k)$ es una pseudo-distancia puesto que no necesariamente debe cumplir con la propiedad de simetría.

- Debe ser estrictamente no negativa (para que tenga sentido como una función de distancia).
- Debe ser estrictamente convexa[2] (para que cualquier mínimo local sea un mínimo absoluto).
- $G(1) = 0$, esto es que la distancia entre pesos iguales es cero.
- $G'(1) = 0$, cuando los pesos son iguales la función debe tener un punto crítico.
- $G''(1) = 1$, ese punto crítico debe corresponder al minimizador.

En resumen, la técnica de calibración induce un nuevo conjunto de pesos w_k que surge de la minimización de una pseudo-distancia $G(\cdot)$ en la muestra que está sujeta a las ecuaciones de calibración. Es decir, que los nuevos pesos deben ser tales que

$$\sum_{k\in S} w_k \mathbf{x}_k = \sum_U \mathbf{x}_k = \mathbf{t_x} \tag{10.3.2}$$

Para resolver este problema de optimización, recurrimos a la técnica de los multiplicadores de Lagrange. De esta manera, la ecuación de Lagrange estará dada por la siguiente expresión

$$\mathcal{L}(w_1,\ldots,w_n,\boldsymbol{\lambda}) = \sum_{k\in S} d_k \frac{G(w_k/d_k)}{q_k} - \boldsymbol{\lambda}'\left(\sum_S w_k\mathbf{x}_k - \mathbf{t_X}\right) \tag{10.3.3}$$

Derivando la ecuación de Lagrange con respecto a w_k e igualando a cero, se tiene

$$\frac{\partial\mathcal{L}}{\partial w_k} = \frac{d_k}{q_k}\frac{g(w_k/d_k)}{d_k} - \boldsymbol{\lambda}'\mathbf{x}_k = 0$$

Donde $g(\omega) = \frac{dG(\omega)}{d\omega}$, y por tanto se llega a que

$$g(w_k/d_k) = q_k\boldsymbol{\lambda}'\mathbf{x}_k$$

En este paso es necesario definir una función $F(\cdot)$, tal que $F(\cdot) = g^{-1}(\cdot)$, es decir $F(g(\omega)) = \omega$, por lo tanto

$$F(g(w_k/d_k)) = F(q_k\boldsymbol{\lambda}'\mathbf{x}_k)$$

Lo que nos guía al valor de los nuevos pesos

$$w_k = d_kF(q_k\boldsymbol{\lambda}'\mathbf{x}_k) \tag{10.3.4}$$

El vector $\boldsymbol{\lambda}$ se obtiene al resolver el siguiente sistema de ecuaciones

$$\sum_{k\in S} \underbrace{d_kF(q_k\boldsymbol{\lambda}'\mathbf{x}_k)}_{w_k}\mathbf{x}_k' = \mathbf{t}_\mathbf{x}' \tag{10.3.5}$$

[2]Una función $G(x)$ es estrictamente convexa sí y sólo sí $G(ax_1 + (1-a)x_2) < aG(x_1) + (1-a)G(x_2)$ para todo $a \in (0,1)$ y todo $x_1 \neq x_2$. Por otro lado, si la segunda derivada de G es positiva en todo su dominio, entonces $G(x)$ es convexa.

10.3.1 Distancias $G(\cdot)$, $g(\cdot)$ y $F(\cdot)$

En general, hay varios tipos de distancias que pueden utilizarse en la construcción de un estimador de calibración. Sin embargo, Deville & Särndal (1992) demuestran que todas ellas guían asintóticamente al mismo estimador. Las pseudo-distancias más utilizadas están dadas en tabla 10.8º. Dependiendo de la escogencia de cada distancia, se obtendrán distintos estimadores de calibración. También es posible fijar dos constantes L y U y restringir el rango de los pesos resultantes w_k al intervalo (L, U). Este método se utiliza para evadir los pesos extremos o negativos, que se pueden eliminar con una buena escogencia de L y U.

En resumen, el proceso para obtener un estimador de calibración es el siguiente:

1. Definir una distancia $G(\cdot)$ y observar los datos y_k y $\mathbf{x}_k$.
2. Resolver (10.3.4) para el vector $\boldsymbol{\lambda}$. En algunos casos esta solución requiere de procedimientos iterativos.
3. Usar $\boldsymbol{\lambda}$ para obtener un estimador del total poblacional de la característica de interés dado por

$$\hat{t}_{y,cal} = \sum_{k \in S} w_k y_k = \sum_{k \in S} d_k F(q_k \boldsymbol{\lambda}' \mathbf{x}_k) y_k \tag{10.3.6}$$

Deville & Särndal (1992) asegura que el estimador $\hat{t}_{y,cal}$ arrojará estimaciones cercanas al total poblacional desconocido de la característica de interés si existe una fuerte relación entre y y $\mathbf{x}$. De hecho, si y estuviera perfectamente explicado por $\mathbf{x}$, la varianza del estimador $\hat{t}_{y,cal}$ sería nula para cada posible muestra.

Tabla 10.8: *Ejemplos de pseudo-distancias para el proceso de calibración.*

Distancia	$G(x)$	$g(x)$	$F(u)$
Ji cuadrado	$\frac{1}{2}(x-1)^2$	$x-1$	$1+u$
Entropía	$x\ln(x) - x + 1$	$\ln(x)$	$\exp(u)$
Hellingster	$2(\sqrt{x}-1)^2$	$2\left(1-\sqrt{\frac{1}{x}}\right)$	$(1+\frac{u}{2})^{-2}$
Entropía inversa	$\ln(\frac{1}{x}) + x - 1$	$1-\frac{1}{x}$	$(1+u)^{-1}$
Ji cuadrado inversa	$\frac{1}{2}\frac{(x-1)^2}{x}$	$\frac{1}{2}\left(1-\frac{1}{x}\right)^2$	$(1+2u)^{-1/2}$

10.4 Algunos casos particulares

Deville & Särndal (1992) examinaron las propiedades estadísticas de $\hat{t}_{y,cal}$ bajo una serie de pseudo-distancias $G(\dot{})$. En esta sección se revisarán algunos casos particulares que arrojan estimadores de calibración, algunos conocidos y otros nuevos.

10.4.1 Método lineal: distancia Ji cuadrado

Este método, quizás el más usado y uno de los más importantes en calibración, se obtiene cuando se escoge la utilizaremos la distancia Ji cuadrado que calcula la distancia, en toda la muestra, de los nuevos pesos w_k a los pesos clásicos d_k como

$$\sum_S d_k G(w_k/d_k) = \frac{1}{2}\sum_S \frac{(w_k - d_k)^2}{d_k}$$

Resultado 10.4.1. *Bajo la distancia Ji cuadrado, y suponiendo que las ponderaciones $q_k = 1/c_k$, el estimador de calibración toma la forma del estimador general de regresión.*

Prueba. De (10.3.3), y utilizando el hecho de que, para este pseudo-distancia, $F(u) = 1 + u$, entonces se tiene que

$$\begin{aligned} w_k &= d_k F(q_k \boldsymbol{\lambda}' \mathbf{x}_k) \\ &= d_k(1 + q_k \boldsymbol{\lambda}' \mathbf{x}_k) \\ &= d_k + d_k q_k \boldsymbol{\lambda}' \mathbf{x}_k \end{aligned}$$

y reemplazando en la ecuación de calibración (10.3.4)

$$\sum_s d_k \mathbf{x}_k' + \sum_s d_k q_k \boldsymbol{\lambda}' \mathbf{x}_k \mathbf{x}_k' = \mathbf{t}_{\mathbf{x}}' \qquad (10.4.1)$$

Al despejar convenientemente, el multiplicador de Lagrange se resuelve como

$$\boldsymbol{\lambda}' = (\mathbf{t}_{\mathbf{x}} - \mathbf{t}_{\mathbf{x}\pi})' \left(\sum_S d_k q_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \qquad (10.4.2)$$

Así, suponiendo que $q_k = 1/c_k$, se llega al estimador de calibración para el total de la característica de interés, puesto que

$$w_k = d_k + d_k(\mathbf{t}_{\mathbf{x}} - \mathbf{t}_{\mathbf{x}\pi})' \mathbf{T}^{-1} q_k \mathbf{x}_k \qquad (10.4.3)$$

donde $\mathbf{T}^{-1}$ está definido en (9.2.13). Entonces, se tiene que

$$\begin{aligned} \hat{t}_{y,cal} &= \sum_S w_k y_k & (10.4.4) \\ &= \sum_s \frac{y_k}{\pi_k} + (\mathbf{t}_{\mathbf{x}} - \hat{\mathbf{t}}_{\mathbf{x}\pi})' \mathbf{T}^{-1} \sum_s \frac{\mathbf{x}_k y_k}{c_k \pi_k} & (10.4.5) \end{aligned}$$

que coincide exactamente con la expresión (9.2.15) que define el estimador general de regresión. ■

El autor recalca que el estimador general de regresión es un caso particular de la familia de estimadores de calibración. Es un error hacer aserciones acerca de los estimadores de calibración basados solamente en la forma funcional del estimador general de regresión (GREG). Aunque es cierto que una gran mayoría de artículos están basados bajo el espíritu del estimador general de regresión, se debe recalcar que la filosofía de un estimador de calibración, aunque no contradice el uso del estimador general de regresión, es bien diferente a la filosofía de éste.

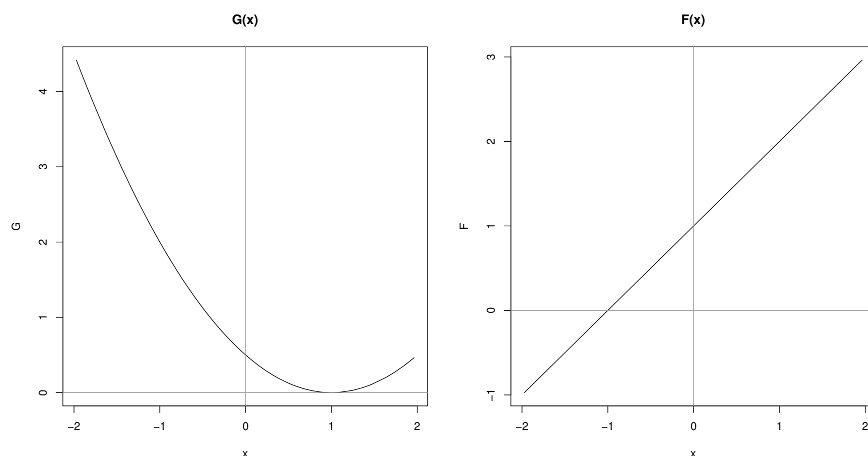


Figura 10.1: *Funciones $G(x)$ y $F(u)$ utilizando la distancia Ji cuadrado.*

Nótese que el estimador general de regresión utiliza un modelo para incorporar la información auxiliar en el proceso de estimación, al igual que los estimadores de calibración, no todos los casos particulares del estimador general de regresión son estimadores de calibración. El espíritu más influyente de los estimadores de calibración no es incorporar un modelo al proceso de estimación sino conseguir un conjunto de pesos w_k. Como Särndal (2007) lo afirma, el concepto de *estimación GREG* y *estimación de calibración* reflejan una clara diferencia de pensamiento. La gran variedad de posibles modelos generan una amplia familia de estimadores tipo GREG. Por otro lado, la escogencia de una distancia en el proceso de calibración generan una amplia familia de estimadores de calibración, cuyo caso particular es la familia de estimadores GREG lineales.

Resultado 10.4.2. *Bajo la distancia Ji cuadrado, y suponiendo que las ponderaciones $q_k = 1/x_k$ y que sólo existe una variable de información auxiliar; es decir $\mathbf{x}_k = x_k$, el estimador de calibración toma la forma del estimador de razón.*

Prueba. Bajo las anteriores condiciones, se tiene que

$$\boldsymbol{\lambda} = \frac{\sum_U x_k}{\sum_S d_k x_k} - 1 = \frac{t_x}{\hat{t}_{x,\pi}} - 1$$

Por tanto

$$w_k = d_k(1 + q_k x_k \boldsymbol{\lambda}) = d_k(1 + \boldsymbol{\lambda}) = d_k \left(\frac{t_x}{\hat{t}_{x,\pi}} \right)$$

Luego, el estimador de calibración toma la forma siguiente

$$\begin{aligned} \hat{t}_{y,cal} &= \sum_S w_k y_k \\ &= \sum_S d_k \frac{t_x}{\hat{t}_{x,\pi}} y_k \\ &= t_x \frac{\hat{t}_{y,\pi}}{\hat{t}_{x,\pi}} = \hat{t}_{y,r} \end{aligned}$$

que coincide con la forma del estimador de razón dada por (9.4.15). ■

10.4.2 Método de raking: distancia de entropía

El método de raking utiliza la distancia de entropía como base de construcción del estimador de calibración. Esta distancia se define como:

$$G(x) = x \log(x) - x + 1$$

Nótese que la distancia, en toda la muestra, de los nuevos pesos w_k a los pesos clásicos d_k como:

$$\sum_S d_k G(w_k/d_k) = \sum_S d_k \left(\frac{w_k}{d_k} \ln \left(\frac{w_k}{d_k} \right) - \frac{w_k}{d_k} + 1 \right)$$

De (10.3.3), y utilizando el hecho de que, para este pseudo-distancia, $F(u) = \exp(u)$, entonces se tiene que

$$\begin{aligned} w_k &= d_k F(q_k \boldsymbol{\lambda}' \mathbf{x}_k) \\ &= d_k \exp(q_k \boldsymbol{\lambda}' \mathbf{x}_k) \end{aligned}$$

y reemplazando en la ecuación de calibración (10.3.4)

$$\sum_s d_k \exp(q_k \boldsymbol{\lambda}' \mathbf{x}_k) \mathbf{x}_k' = \mathbf{t}_{\mathbf{x}}' \qquad (10.4.6)$$

El anterior sistema debe ser resuelto para $\boldsymbol{\lambda}$ (que es un vector columna de multiplicadores de Lagrange). Después de que $\boldsymbol{\lambda}$ sea determinado, se calculan los pesos calibrados como $w_k = d_k \exp(q_k \boldsymbol{\lambda}' \mathbf{x}_k)$ y se obtiene el estimador de calibración para el total poblacional de la característica de interés, definido como:

$$\hat{t}_{y,cal} = \sum_S w_k y_k = \sum_S d_k \exp(q_k \boldsymbol{\lambda}' \mathbf{x}_k) y_k \quad (10.4.7)$$

¿Qué interpretación teórico-práctica tiene que algún w_k resulte negativo? Un aspecto realmente importante de este método de raking es que induce pesos w_k que son estrictamente positivos, lo cual no sucede con el método lineal.

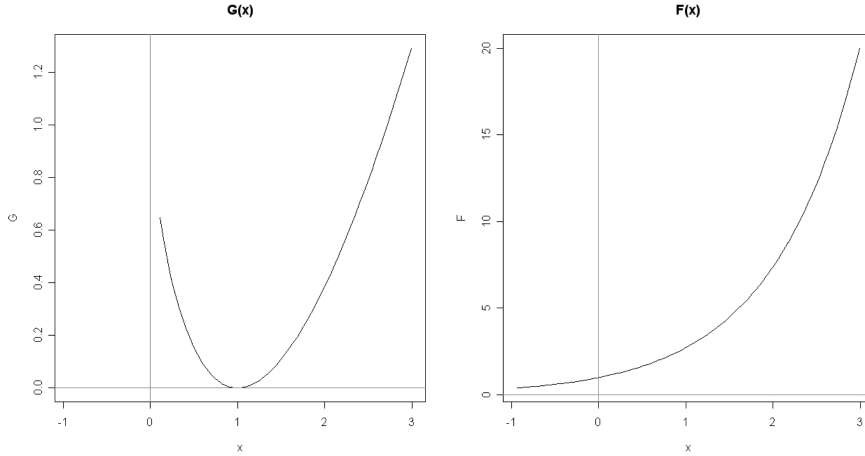


Figura 10.2: *Funciones $G(x)$ y $F(u)$ utilizando la distancia de Entropía.*

Aspectos computacionales para el cálculo de $\boldsymbol{\lambda}$

Para calcular el estimador de calibración dado por (10.4.7), es necesario resolver el sistema de ecuaciones (10.4.6) para $\boldsymbol{\lambda}$. En Deville & Särndal (1992), se demuestra que una solución general pede ser obtenida usando el método iterativo de Newthon-Raphson. Nótese que el sistema de ecuaciones de calibración puede ser re-escrito como una función ϕ en términos de $\boldsymbol{\lambda}$, así:

$$\phi(\boldsymbol{\lambda}) = \sum_S d_k \exp(q_k \boldsymbol{\lambda}' \mathbf{x}_k) \mathbf{x}_k' - \mathbf{t}_{\mathbf{x}}'$$

Nótese que la derivada de esta función con respecto a $\boldsymbol{\lambda}$ está dada por:

$$\phi'(\boldsymbol{\lambda}) = \frac{\partial \phi(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} = \sum_S d_k \exp(q_k \boldsymbol{\lambda}' \mathbf{x}_k) \mathbf{x}_k' \mathbf{x}_k$$

para algún vector $\boldsymbol{\lambda}$. Entonces, de acuerdo con el método de Newton-Raphson, una solución estaría dada por la iteración hasta convergencia de la siguiente expresión

$$\boldsymbol{\lambda}^{(a+1)} = \boldsymbol{\lambda}^{(a)} - \left[\phi'\left(\boldsymbol{\lambda}^{(a)}\right)\right]^{-1} \phi(\boldsymbol{\lambda}^{(a)}) \quad (10.4.8)$$

Nótese que el procedimiento converge cuando la diferencia entre $\boldsymbol{\lambda}^{(a+1)}$ y $\boldsymbol{\lambda}^{(a)}$ sea menor que una tolerancia fijada de antemano. Además, se debe tener en cuenta que $\boldsymbol{\lambda}^{(0)} = \mathbf{0}$.

Resultado 10.4.3. *Bajo el método de Newton-Raphson, la primera iteración del algoritmo da como resultado la solución para* $\boldsymbol{\lambda}$ *cuando se utilizaba la distancia Ji cuadrado. Es decir,*

$$\boldsymbol{\lambda}^{(1)} = \hat{\mathbf{T}}^{-1}[\mathbf{t_x} - \hat{\mathbf{t}}_{\mathbf{x}\pi}] \tag{10.4.9}$$

Prueba.

$$\begin{aligned}
\boldsymbol{\lambda}^{(1)} &= \boldsymbol{\lambda}^{(0)} - \left[\phi'\left(\boldsymbol{\lambda}^{(0)}\right)\right]^{-1} \phi(\boldsymbol{\lambda}^{(0)})\\
&= -\left[\sum_s d_k \exp(\mathbf{x}_k\boldsymbol{\lambda}^{(0)})\mathbf{x}'_k\mathbf{x}_k\right]^{-1}\left[\sum_s d_k \exp(\mathbf{x}_k\boldsymbol{\lambda}^{(0)})\mathbf{x}_k - \mathbf{t}_x\right]\\
&= -\left[\sum_s d_k\mathbf{x}'_k\mathbf{x}_k\right]^{-1}\left[\sum_s d_k\mathbf{x}_k - \mathbf{t}_x\right]\\
&= -\hat{\mathbf{T}}^{-1}[\hat{\mathbf{t}}_{\mathbf{x}\pi} - \mathbf{t_x}] = \hat{\mathbf{T}}^{-1}(\mathbf{t_x} - \hat{\mathbf{t}}_{\mathbf{x}\pi})
\end{aligned}$$

que coincide con la solución para $\boldsymbol{\lambda}$ dada por la expresión (10.4.2). ■

Del anterior resultado se tiene que el estimador de calibración en la primera iteración estaría dado por

$$\hat{t}^{(1)}_{y,cal} = \sum_S d_k \exp(q_k(\mathbf{t_x} - \hat{\mathbf{t}}_{\mathbf{x}\pi})'\hat{\mathbf{T}}^{-1}\mathbf{x}_k)y_k \tag{10.4.10}$$

Programación del estimador con R

En esta sección se dan las ideas básicas para la programación computacional de un estimador de calibración basado en el método de raking para el caso en que se utiliza una sola variable de información auxiliar. Nótese que en el cálculo del vector $\boldsymbol{\lambda}$, cuya expresión está dada por la ecuación (10.4.8), están involucradas las funciones ϕ y ϕ'. La programación computacional de esta técnica de los estimadores de calibración puede ser fácilmente implementada en cuatro sencillos pasos. A saber:

1. Programar la función ϕ
2. Programar la función ϕ'
3. Utilizar las anteriores expresiones para realizar el cálculo del vector $\boldsymbol{\lambda}$
4. Iterar hasta convergencia

En la programación de la función ϕ intervienen cuatro *objetos computacionales* los cuales son el vector $d_k = (1/\pi_1, \ldots, 1/\pi_k, \ldots, 1/\pi_n)$, el vector $\boldsymbol{\lambda}$, el vector de valores auxiliares para cada elemento incluido en la muestra, dado por

$\mathbf{x}_k = (x_{1k}, x_{2k}, \ldots, x_{pk})$ y el vector de totales poblacionales de las variables de información auxiliar $\mathbf{t_x}$. De esta manera, el siguiente código crea una función que permite el cálculo de la función ϕ.

```
> Fi <- function(dk, l, x, tx){
+ e <- matrix(0,n,1)
+ for(k in 1:n) {
+               e[k] <- exp(x[k]*l)
+               }
+ res <- sum(dk*e*x)-tx
+ res
+ }
```

Por otra parte, en la programación de la función ϕ' intervienen sólo tres *objetos computacionales* que también estuvieron involucrados en la programación de la función ϕ. La razón de lo anterior es porque ϕ' es la derivada de ϕ. Estos elementos son $d_k = (1/\pi_1, \ldots, 1/\pi_k, \ldots, 1/\pi_n)$, $\boldsymbol{\lambda}$ y $\mathbf{x}_k = (x_{1k}, x_{2k}, \ldots, x_{pk})$. Luego, el siguiente código crea una función que permite el cálculo de la función ϕ'.

```
> Fiprima <- function(dk, l, x){
+ e <- matrix(0,n,1)
+ for(k in 1:n) {
+               e[k] <- exp(x[k]*l)
+               }
+ res <- sum(dk*e*x*x)
+ res
+ }
```

Simultáneamente, se debe crear una función que calcule el estimador de calibración. En esta función intervienen cuatro *objetos computacionales* que son: $d_k = (1/\pi_1, \ldots, 1/\pi_k, \ldots, 1/\pi_n)$, $\boldsymbol{\lambda}$, $\mathbf{x}_k = (x_{1k}, x_{2k}, \ldots, x_{pk})$ y por último el vector de valores de la característica de interés para los elementos de la muestra $y_k = (y_1, y_2, \ldots, y_n)$.

```
> Cal <- function(dk, l, x, y){
+ w <- matrix(0,n,1)
+ for(k in 1:n) {
+               w[k] <- exp(x[k]*l)
+               }
+ res <- sum(dk*w*y)
+ res
+ }
```

Por supuesto, los anteriores códigos no funcionan por sí solos. Nótese que las anteriores funciones tienen al elemento computacional $\boldsymbol{\lambda}$ en común; sin embargo, este elemento no existe aún y debe ser calculado con métodos iterativos como el de Newton-Raphson. Estas funciones deben ser ensambladas por una función que las recoja y que sea capaz de realizar el cálculo final del estimador de calibración.

En primer lugar se debe fijar una tolerancia deseada, en este caso la tolerancia está dada por 0.000001. Esto quiere decir que el proceso iterativo se detiene cuando suceda que $|\boldsymbol{\lambda}^{(a+1)} - \boldsymbol{\lambda}^{(a)}| < -0.000001$. Sin embargo, si esta condición no se satisface, entonces el proceso sigue iterándose repetitivamente. Cuando el proceso converge, entonces es posible utilizar las funciones que se declararon anteriormente y así calcular el valor de la estimación.

```
> tol <- 0.000001
> l <- 0
> l.k <- 4

> while(abs(l-l.k)>tol){
+                      l.k <- l-Fi(l,xs,sum(xu))/Fiprima(l,xs)
+                      l <- l.k
+                      }

> tcal <- Cal(l.k, xs, ys)
```

Los anteriores códigos de programación pretenden ser una guía para el estudiante y no se declaran como la única alternativa de lógica computacional.

Nótese, sin embargo, que aunque el método de Raking posee la característica de que los pesos no son negativos, como suele suceder cuando se utiliza el método lineal, éstos pueden ser muy variables. Para resolver este inconveniente, Deville & Särndal (1992) proponen los métodos logístico y lineal truncado. Éstas técnicas surgen motivadas por el deseo de restringir el rango de variación de los nuevos pesos de calibración sin alterar demasiado el estimador de calibración. En la práctica, el estadístico desea evadir los pesos extremos; en la siguientes secciones se muestra cómo estos pueden ser eliminados.

10.4.3 Método logístico

Conocido comúnmente como el método de calibración Logit (L, U). Este método fija dos constantes L y U tales que $L < 1 < U$. De esta forma se define la siguiente cantidad

$$A = \frac{(U - L)}{(1 - L)(U - 1)}$$

Luego, se define la siguiente función

$$G(x) = \begin{cases} \frac{1}{A}\left[(x - L)\lg\frac{x-L}{1-L} + (U - x)\lg\frac{U-x}{U-1}\right] & \text{si } L < x < U \\ \\ \infty & \text{en otro caso} \end{cases} \tag{10.4.11}$$

La correspondiente función F está dada por

$$F(u) = \frac{L(U - 1) + U(1 - L)\exp(Au)}{U - 1 + (1 - L)\exp(Au)} \tag{10.4.12}$$

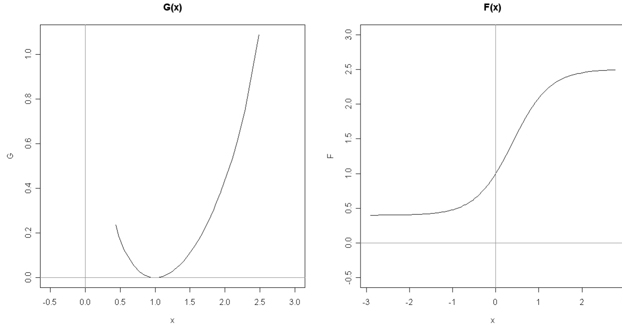


Figura 10.3: *Funciones $G(x)$ y $F(u)$ utilizando el método logístico con $L = 0.4$ y $U = 2.5$la distancia de Entropía.*

La anterior función toma valores restringidos al intervalo (L, U) puesto que $F(-\infty) = L$ y $F(\infty) = U$. Por lo tanto los nuevos pesos de calibración están siempre en el intervalo $[Ld_k, Ud_k]$.

10.4.4 Método truncado lineal

Para restringir el intervalo de soluciones de los pesos de calibración es posible utilizar la misma función lineal pero restringida a dos valores L y U, tales que $L < 1 < U$. De esta forma,

$$G(x) = \begin{cases} \frac{1}{2}(x-1)^2 & \text{si } L < x < U \\ \\ \infty & \text{en otro caso} \end{cases} \qquad (10.4.13)$$

De esta manera, la correspondiente función F, está dada por

$$F(u) = \begin{cases} 1+u & \text{si } u \in [L-1, U-1] \\ \\ L & \text{si } u < L-1 \\ \\ U & \text{si } u > U-1 \\ \end{cases} \qquad (10.4.14)$$

Así, los nuevos pesos de calibración están siempre en el intervalo $[Ld_k, Ud_k]$.

10.5 Calibración y Post-estratificación

Deville, Särndal & Sautory (1993) derivaron en primer lugar el estimador de calibración y luego explicaron el estimador de post-estratificación y el estimador de

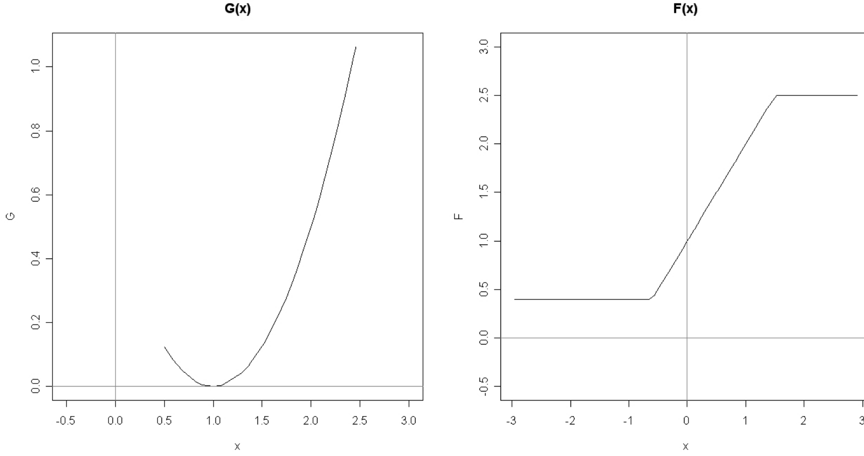


Figura 10.4: *Funciones $G(x)$ y $F(u)$ utilizando el método truncado lineal con $L = 0.4$ y $U = 2.5$la distancia de Entropía.*

Raking (bajo el algoritmo IPFP) como casos particulares del método de calibración bajo distintas distancias. En esta sección se dan las bases estadísticas para la construcción de estos estimadores.

10.5.1 Post-estratificación

Un caso especial muy importante de los estimadores de calibración corresponde al estimador de post-estratificación completa[3]. En este caso el número de variables de información auxiliar es igual al número de post-estratos que particionan la población. Este proceso supone la partición en G grupos de la población finita. Así que $U = (U_1, U_2, \ldots, U_G)$. Se asume que la característica de interés está relacionada con G vectores o variables dummy que toman el valor uno si el elemento pertenece al subgrupo U_g $(g = 1, \ldots, G)$ o cero si el elemento no pertenece al grupo. Así que $p = G$, $\mathbf{x}_k = \mathbf{d}_k = \underbrace{(0, 0, \ldots, 1, \ldots, 0, 0)}_{G \texttt{ grupos}}'$ y $q_k = 1$ para todo $k \in U$.

Bajo la anterior formulación tenemos que el vector λ toma la siguiente forma

$$\boldsymbol{\lambda}' = (\lambda_1, \ldots, \lambda_g, \ldots, \lambda_G) \tag{10.5.1}$$

y cada entrada del vector de información auxiliar para el k-ésimo elemento está dada por

$$x_{kg} = \begin{cases} 1 & \text{si } k \in U_g \\ \\ 0 & \text{en otro caso} \end{cases} \tag{10.5.2}$$

[3]El término post-estratificación completa se usa cuando los totales internos de la tabla de contingencia son conocidos y se usan para el proceso de calibración.

Nótese que

$$\mathbf{t_x} = \sum_{k\in U} \mathbf{x}'_k = (N_1, \ldots, N_g, \ldots, N_G), \qquad (10.5.3)$$

donde N_g corresponde al total de elementos pertenecientes al subgrupo poblacional U_g.

Resultado 10.5.1. *Los pesos de calibración para el caso de post-estratificación están dados por*

$$w_k = d_k \frac{N_g}{\hat{N}_{g,\pi}} \qquad g = 1, \ldots, G \qquad (10.5.4)$$

y son invariantes a la escogencia de cualquier distancia.

Prueba. La construcción del estimador de calibración para este esquema particular es como sigue. En primer lugar, nótese que si el k-ésimo elemento pertenece al subgrupo U_g, entonces

$$\boldsymbol{\lambda}'\mathbf{x}_k = \lambda_g \qquad (10.5.5)$$

Por tanto la restricción de calibración dada por

$$\sum_{k\in S} d_k F(\boldsymbol{\lambda}'\mathbf{x}_k)\mathbf{x}'_k = \mathbf{t}'_\mathbf{x} \qquad (10.5.6)$$

puede ser re-escrita como

$$\sum_{k\in U_g} d_k F(\lambda_g) = N_g \qquad g = 1, \ldots, G \qquad (10.5.7)$$

Por tanto, despejando la anterior ecuación, se tiene finalmente que

$$F(\lambda_g) = \frac{N_g}{\sum_{k\in U_g} d_k} = \frac{N_g}{\hat{N}_{g,\pi}} \qquad g = 1, \ldots, G \qquad (10.5.8)$$

Luego, de (10.3.3) los pesos de calibración están dados por

$$w_k = d_k \frac{N_g}{\hat{N}_{g,\pi}} \qquad g = 1, \ldots, G \qquad (10.5.9)$$

Nótese que en la construcción de los pesos de calibración no importó la escogencia de la distancia. ■

Por tanto el estimador de calibración está dado por

$$\hat{t}_{y,cal} = \sum_{k\in S} w_k y_k$$
$$= \sum_{g=1}^{G} \sum_{k\in S_g} \frac{N_g}{\hat{N}_{g,\pi}} \frac{y_k}{\pi_k}$$

que equivale al estimador de post-estratificación.

10.5.2 Raking

Si Deming hubiese dado cuenta de los estimadores de calibración cuando se usa la distancia multiplicativa como marco de referencia, hubiera estado muy contento al darse cuenta de que su método pudo ser generalizado e incluido en el contenido de la ciencia estadística. Al principio, el IPFP se usó de manera totalmente pragmática, simplemente se trataba de realizar un ajuste para que las estimaciones internas de la tabla de contingencia calibraran los totales conocidos. Bajo este marco de referencia, el IPFP era criticado por ser un método matemático y no estadístico cuyos resultados no tenían en cuenta el diseño de muestro que se había usado para la recolección de la información. Como se verá en esta sección, el estimador de calibración que apunta a la estimación de las celdas internas en tablas de contingencia es equivalente al resultante del método IPFP. De hecho, el método IPFP es un caso particular de este escenario que se conoce con el nombre de Raking.

Como caso particular se considera la estimación de una tabla de contingencia a dos vías con calibración sobre los totales marginales. Por lo anterior, la partición de la población sigue el patrón de la siguiente tabla.

Tabla 10.9: *Partición de la población.*

U_{11}	$\cdots$	U_{1g}	$\cdots$	U_{1G}	$U_{1\cdot}$
$\vdots$		$\vdots$		$\vdots$	$\vdots$
U_{h1}	$\cdots$	U_{hg}	$\cdots$	U_{hG}	$U_{h\cdot}$
$\vdots$		$\vdots$		$\vdots$	$\vdots$
U_{H1}	$\cdots$	U_{Hg}	$\cdots$	U_{HG}	$U_{H\cdot}$
$U_{\cdot 1}$	$\cdots$	$U_{\cdot g}$	$\cdots$	$U_{\cdot G}$	U

Se supone que $q_k = 1$ para todo $k \in U$ y $\mathbf{x}_k = (\mathbf{d}'_{1k}, \mathbf{d}'_{1k})$, donde $\mathbf{d}_{1k}$ es un vector de H variables dummy denotando a cuál post-estrato pertenece el k-ésimo elemento y $\mathbf{d}_{2k}$ es un vector de G variables dummy denotando a cuál post-estrato pertenece el k-ésimo elemento. Nótese que

$$\mathbf{t_x} = \sum_{k\in U} \mathbf{x}'_k = (N_{1\cdot}, \ldots, N_{h\cdot}, \ldots, N_{H\cdot}, N_{\cdot 1}, \ldots, N_{\cdot g}, \ldots, N_{\cdot G}) \qquad (10.5.10)$$

Sea $\mathbf{u} = (u_1, \ldots, u_H)'$ un vector de orden H y $\mathbf{v} = (v_1, \ldots, v_G)'$ un vector de orden G. Definiendo $\boldsymbol{\lambda}' = (\mathbf{u}', \mathbf{v}')$, se tiene que si el k-ésimo elemento pertenece a la celda U_{hg}, entonces

$$F(q_k\boldsymbol{\lambda}'\mathbf{x}_k) = F(u_h + v_g) \tag{10.5.11}$$

Por tanto las ecuaciones de calibración (10.5.6) pueden ser escritas como el siguiente sistema de ecuaciones

$$\sum_{h=1}^{H} \hat{N}_{hg,\pi} F(u_h + v_g) = N_{\cdot g} \qquad g = 1, \ldots, G \tag{10.5.12}$$

$$\sum_{g=1}^{G} \hat{N}_{hg,\pi} F(u_h + v_g) = N_{h\cdot} \qquad h = 1, \ldots, H \tag{10.5.13}$$

donde $\hat{N}_{hg,\pi}$ corresponde al estimador de Horvitz-Thompson de N_{hg}. Si se utiliza la distancia de entropía, se tiene que

$$F(u_h + v_g) = \exp(u_h + v_g) = \exp(u_h)\exp(v_g) \tag{10.5.14}$$

Por tanto el sistema de ecuaciones dado por (10.5.12) y (10.5.13) toma la siguiente forma

$$\exp(u_h) = \frac{N_{\cdot g}}{\sum_{g=1}^{G} \hat{N}_{hg}\exp(v_g)} \qquad h = 1, \ldots, H \tag{10.5.15}$$

$$\exp(v_g) = \frac{N_{h\cdot}}{\sum_{h=1}^{H} \hat{N}_{hg}\exp(v_h)} \qquad g = 1, \ldots, G \tag{10.5.16}$$

Una solución para el anterior sistema de ecuaciones se obtiene al iterar hasta convergencia el algoritmo IPFP como sigue.

1. Fijar $\exp(v_g) = 1$ y calcular $\exp(u_h)$ en (10.5.15)
2. Luego insertar este valor de $\exp(u_h)$ en (10.5.16) y calcular un nuevo valor de $\exp(v_g)$
3. Iterar hasta convergencia

Después de que el algoritmo ha finalizado, el estimador de calibración para el total de la celda U_{hg} está dado por

$$\hat{N}_{hg,cal} = \hat{N}_{hg,\pi}\exp(u_h + v_g) = \hat{N}_{hg,\pi}\exp(u_h)\exp(v_g) \tag{10.5.17}$$

y los nuevos pesos calibrados son $w_k = d_k \exp(u_h + v_g) = d_k \exp(u_h)\exp(v_g)$ si el k-ésimo elemento pertenece a la celda U_{hg}.

10.6 Varianza de los estimadores de calibración

Cerramos este capítulo con una importante propiedad de los estimadores de calibración.

Resultado 10.6.1. *El estimador de calibración es asintóticamente equivalente al estimador general de regresión bajo las siguientes condiciones de regularidad:*

1. $\lim \frac{\mathbf{t_x}}{N}$ *existe*
2. $\frac{\hat{\mathbf{t}}_{\mathbf{x}\pi} - \mathbf{t_x}}{N} \to \mathbf{0}$ *en probabilidad*[4]
3. $\sqrt{n}\frac{\hat{\mathbf{t}}_{\mathbf{x}\pi} - \mathbf{t_x}}{N}$ *converge en distribución a la normal multivariante* $N(\mathbf{0}, \mathbf{A})$

Prueba. La demostración del anterior resultado se sale del alcance de este libro. Sin embargo, el lector interesado puede consultar en Deville & Särndal (1992). ■

En particular, bajo el anterior resultado, el estimador de calibración comparte las mismas esperanzas asintóticas y las mismas varianzas asintóticas que el estimador general de regresión. Esto puede ser visto mediante el siguiente argumento heurístico:

- Se asume que para tamaños de muestra grandes el estimador de Horvitz-Thompson, $\hat{\mathbf{t}}_{\mathbf{x}\pi}$, es cercano al total poblacional de las características de información auxiliar, $\mathbf{t_x}$. Lo anterior se tiene puesto que $\hat{\mathbf{t}}_{\mathbf{x}\pi}$ es un estimador consistente para $\mathbf{t_x}$.
- Entonces, siguiendo la ecuación (10.3.4), el valor de $F(\cdot)$ debería ser cercano a uno y el valor de $\boldsymbol{\lambda}$ debería ser cercano a $\mathbf{0}$
- Sin embargo, por la construcción de las funciones $F(\cdot)$ y dado que $F(0) = F'(0) = 1$, entonces todas las funciones $F(\cdot)$ deberían tener el mismo comportamiento en la vecindad de 0.
- Por tanto, todas las funciones $F(\cdot)$ pueden ser aproximadas mediante la función $F(u) = u + 1$.
- Es decir, la misma función que corresponde al estimador general de regresión.

Resultado 10.6.2. *La varianza aproximada y la estimación de la varianza del estimador de calibración está dada por.*

$$AV(\hat{t}_{y,cal}) = \sum\sum_U \Delta_{kl}(d_k E_k)(d_l E_l) \tag{10.6.1}$$

$$\widehat{Var}(\hat{t}_{y,cal}) = \sum\sum_S \frac{\Delta_{kl}}{\pi_{kl}}(w_k e_k)(w_l e_l) \tag{10.6.2}$$

respectivamente. Donde $E_k = y_k - \mathbf{x}_k'\mathbf{B}$ *y* $\mathbf{B}$ *satisface las ecuaciones normales en la construcción del estimador de regresión. También* $e_k = y_k - \mathbf{x}_k'\hat{\mathbf{B}}$ *y* $\hat{\mathbf{B}}$ *es un estimador de* $\mathbf{B}$.

[4]El marco de referencia de esta medida de probabilidad está dado por el diseño muestral que se utilizó en la estrategia de muestreo.

10.7 Marco y Lucy

Volviendo con el ejercicio práctico de estimación, suponga que el gobierno desea obtener una estimación del total de impuestos que el sector industrial aportó en el último año fiscal. Estas estimaciones se requiere que sean muy precisas puesto que con base en estos resultados se replanteará una parte del presupuesto nacional.

En esta ocasión, el gobierno pone a disposición del estadístico un marco de muestreo que incluye la identificación y ubicación de todas las empresas pertenecientes al sector industrial. Además de esto, el gobierno tiene la disponibilidad de conocimiento del total poblacional de dos características de información auxiliar; a saber, el total poblacional de la variable `Employees` correspondiente a 151950, el total poblacional de la variable `Income` correspondiente a 1035217 y, por supuesto, el total poblacional del número de empresas del sector industrial correspondiente a 2396.

Bajo el anterior esquema, se planearon varias estrategias de muestreo que manejaban un diseño aleatorio simple de 400 empresas y estimadores de calibración bajo varias distancias. Para la selección de tal muestra se utilizó el siguiente código computacional

```
> data(Lucy)
> attach(Lucy)
> N <- dim(Lucy)[1]
> n <- 400 > sam <- sample(N,n)
> muestra <- Lucy[sam,]
> attach(muestra)
```

Una vez que la muestra fue seleccionada se utilizó el paquete `sampling` del software R para calcular lo estimadores de calibración. En particular se utilizó la función `calib` que calcula los pesos w_k del estimador de calibración. Esta función cuenta con varios argumentos; entre ellos están lo siguientes: `Xs`, la matriz que contiene los valores de las características de información auxiliar para los individuos incluidos en la muestra, `d`, correspondiente al inverso de los pesos de las probabilidades de inclusión de los elementos en la muestra, `tx`, que corresponde al total poblacional de las variables de calibración, `method` que incluye cuatro posibles distancias que son la distancia Ji cuadrado cuyo acepción en la función `calib` está dada por `method=''linear''`, la distancia de entropía cuya acepción en la función `calib` está dada por `method=''raking''` y los métodos logístico y truncado cuyas acepciones en la función `calib` están dadas por `method=''logit''` y por `method=''truncated''`, respectivamente.

Se calcularon las estimaciones de calibración usando los cuatro métodos y el código utilizado se muestra a continuación.

```
> library(sampling)
> ys <- data.frame(Income, Employees, Taxes)
> Xs <- cbind(1,Income, Employees)
> piks <- rep(n/N,times=n)
> tx <- c(N,1035217,151950)
> w1 <- calib(Xs, d=1/piks, tx, method="linear")
```

```
> w2 <- calib(Xs, d=1/piks, tx, method="raking")
> w3 <- calib(Xs, d=1/piks, tx, method="logit",
+                                    bounds=c(0.75,1.2))
> w4 <- calib(Xs, d=1/piks, tx, method="truncated",
+                                    bounds=c(0.75,1.2))
```

La función `calib` solamente calcula los pesos que intervienen en las ecuaciones de calibración. Para calcular las estimación final del total de la característica de interés `Taxes` se debe proceder a multiplicar las cantidades pertinentes. De esta manera, el siguiente código se utilizó para el cálculo de las cuatro estimaciones.

```
> tcal1 <- t(w1/piks)%*%as.matrix(ys)
> tcal2 <- t(w2/piks)%*%as.matrix(ys)
> tcal3 <- t(w3/piks)%*%as.matrix(ys)
> tcal4 <- t(w4/piks)%*%as.matrix(ys)
> tcal1
      Income Employees    Taxes
[1,] 1035217    151950  28729.50
> tcal2
      Income Employees    Taxes
[1,] 1035217    151950  28851.88
> tcal3
      Income Employees    Taxes
[1,] 1035218  151950.1  29018.58
> tcal4
      Income Employees    Taxes
[1,] 1035217    151950  28994.22
```

La tabla 10.10. muestra el comportamiento de los estimadores de calibración del total de los impuestos declarados por las empresas de grupo industrial. Nótese que la diferencia no es grande entre ellos. Por otra parte, la ganancia en eficiencia relativa es alta comparada con la estrategia de muestreo simple.

Tabla 10.10: *Estimaciones realizadas con diferentes métodos de calibración para el total de la características de interés Impuesto.*

Método	Total poblacional	Total estimado	Desv. %
Lineal	28654	28729	0.26
Raking	28654	28851	0.68
Logístico	28654	29018	1.27
Truncado	28654	28994	1.18

10.8 Discusión

Särndal (2007) afirma que la definición del enfoque de calibración para la estimación de totales en poblaciones finitas sigue los siguientes procesos:

1. Calcular nuevos pesos que incorporen información auxiliar específica y que están restringidos a la ecuación de calibración.

2. Utilizar estos nuevos pesos para la construcción de estimadores lineales.

3. Obtener estimaciones aproximadamente insesgadas en presencia de no respuesta y otros errores no muestrales.

Al mismo tiempo, Särndal (2007) concluye que existen seis ideas sobre las cuales vale la pena profundizar un poco más. A continuación se exponen estos criterios que algunos estadísticos han usado para enfatizar el uso práctico de los estimadores de calibración:

- **Como un método de ponderación lineal:** la calibración tiene un vínculo íntimo con la práctica. La fijación con métodos de ponderación de las agencias que manejan las estadísticas oficiales es una poderosa costumbre en la práctica que empezó con la ponderación de unidades mediante el inverso de su probabilidad de inclusión y siguió con las ponderaciones surgidas del enfoque de post-estratificación. Las ponderaciones de calibración extienden las anteriores ideas. La calibración es nueva como término en el muestreo (casi 15 años) pero no es nueva como una técnica para producir ponderaciones, por ejemplo, el muestreo por cuotas es una forma de muestreo no probabilístico que induce estimaciones calibradas con los totales demográficos de la población de estudio. La ponderación de los valores observados de las características de interés fue un tópico muy importante antes que el término calibración comenzara a ser popular. Algunos autores derivaron estas ponderaciones con el argumento que deberían diferir de la manera más mínima posible de los pesos originales. Otros autores encontraron las ponderaciones al reconocer que un estimador de regresión lineal podría ser escrito como una suma ponderada de los valores de la característica de interés. De allí surgieron términos tales como ponderación de muestreo, ponderación de regresión y ponderación de caso.

- **Como una forma sistemática para utilizar la información auxiliar:** la calibración provee una forma sistemática para involucrar la información auxiliar. En la mayoría de aplicaciones práctica la calibración provee un enfoque simple para incorporar esta información dentro de la etapa de estimación. La información auxiliar fue usada para mejorar la precisión de los estimativos mucho antes que el término calibración fuera popular. Existen cientos de artículos que fueron escritos con este propósito en mente. Hoy en día la calibración ofrece un camino para incorporar esta información auxiliar. Por ejemplo la calibración puede ser usada efectivamente en encuestas donde la información auxiliar está disponible en diferentes niveles. Al realizar un muestreo en dos etapas la información auxiliar puede existir para las unidades de la primera etapa (los conglomerados) y puede existir otra información para las unidades de la segunda etapa (elementos o conglomerados).

- **Como un enfoque para conseguir consistencia:** en algunas ocasiones el término calibración se refiere a una forma de conseguir estimativos consistentes[5]. Las ecuaciones de calibración imponen la característica de consistencia sobre el vector de ponderaciones; así que, cuando éste se aplica a las variables

[5]En este apartado la palabra consistente se da en el sentido de la consistencia con los totales de la información auxiliar.

auxiliares el resultado será consistente con los totales de estas variables. Un deseo de promover la credibilidad en las estadísticas oficiales es una razón para que las entidades busquen la consistencia. Cuando la motivación primaria para la calibración no es la concordancia con los totales de la información auxiliar sino el reducir la varianza y el sesgo debido a la ausencia de respuesta entonces el vector de ponderaciones se dice balanceado.

- **Como excusa de transparencia y conveniencia:** el enfoque de calibración ha ganado popularidad en las aplicaciones reales debido a que las estimaciones resultantes son fáciles de interpretar y de motivar puesto que están directamente relacionadas a los pesos inducidos por el diseño de muestreo. La calibración sobre los totales conocidos brinda al usuario una forma natural y transparente de estimación. El usuario que entiende la ponderación muestral aprecia el método de calibración puesto que modifica sutilmente los pesos originales, pero al mismo tiempo respeta los totales de la información auxiliar y mantiene el sesgo despreciable. Existe otra ventaja que es apreciada por los usuarios, en la mayoría de aplicaciones, la calibración induce un único vector de ponderaciones aplicable a todas las variables involucradas en el estudio. Esta última razón hace que este método sea muy apetecido en las entidades oficiales que manejan encuestas muy extensas.

- **En combinación cono otros términos:** Algunos autores usan la palabra calibración en combinación con otros términos para describir varias direcciones de pensamientos, entre esta proliferación de términos están: calibración modelo, calibración G, calibración armonizada, calibración a un nivel más alto, calibración de regresión, calibración no lineal, calibración super-generalizada, calibración de modelos de redes neuronales y calibración basada en modelos locales polinomiales, entre otras. La calibración juega un rol significativo en los métodos de muestreo indirectos (ver capítulo 12). Este término también ha sido usado, aunque en un espíritu diferente, en conceptos tales como imputación calibrada y calibración sesgada.

- **Como una nueva dirección de pensamiento:** si la calibración representa un nuevo enfoque demarcado claramente de sus predecesores, entonces es tiempo de hacer la pregunta: ¿La calibración generaliza las teorías anteriores? ¿La calibración da mejores respuestas a las preguntas de importancia, que los enfoques de estimación anteriores? En la práctica el estadístico encuentra algunos pormenores tales como ausencia de respuestas, deficiencias del marco muestral y errores de medición. Es cierto que algunos procesos como la imputación y la reponderación para no respuestas son ampliamente difundidos y usados en la práctica. Sin embargo queda un sinsabor al utilizar estos métodos pues no están enmarcados dentro de una teoría exhaustiva de inferencias en poblaciones finitas. La mayoría de artículos teóricos tratan con la estimación de parámetros bajo un mundo ideal, que no existe en la práctica, donde la ausencia de respuesta y otros errores no muestrales están ausentes.

10.9 Estimadores óptimos de calibración

Como lo afirma Wu (2003) existen dos variantes en la construcción de un estimador de calibración: una está dada por la escogencia de la distancia y la otra por el conjunto de ecuaciones de calibración[6] en áreas como la demografía existe la costumbre de calibrar sobre muchas variables, para que se logre estimar con varianza nula los totales conocidos de las variables auxiliares, sin importar que el estimador resultante pueda perder eficiencia. En estos términos, sería mejor utilizar la menor cantidad de ecuaciones de calibración para no estropear el buen comportamiento del estimador. La pregunta que debe plantearse el investigador es ¿cuál es la mejor ecuación de calibración que se debe usar en la construcción de un estimador de este tipo?

Si $u_k = u(\mathbf{x}_k)$, donde $u(\cdot)$ es una función de valor real, entonces una nueva forma de construir un estimador de calibración estaría dada por la consecución de unos pesos w_k restringidos[7] a

$$\sum_{k \in S} w_k u(\mathbf{x}_k) = \sum_{k \in U} u(\mathbf{x}_k)$$

Por tanto, la pregunta se torna más diáfana y se convierte en ¿cuál función $u(\cdot)$ hace al estimador $\hat{t}_{ycal}$ más eficiente? Ahora, es bien sabido que bajo la inferencia basada en el diseño de muestreo, no existe un estimador insesgado de mínima varianza uniformemente (Cassel, Särndal & Wretman 1976*a*). Sin embargo, es posible obtener un estimador óptimo bajo la inferencia asistida por modelos de super-población. La respuesta a estas preguntas está dada por la propuesta de Wu (2003) que construyó un estimador óptimo de calibración suponiendo que las respuestas de y_k pueden ser vistas como realizaciones del siguiente modelo de super-población semi-paramétrica

$$E_\xi(y_k|\mathbf{x}_k) = \mu(\mathbf{x}_k, \boldsymbol{\theta}), \quad Var_\xi(y_k|\mathbf{x}_k) = [\nu(\mathbf{x}_k)]^2\sigma^2 \quad , \tag{10.9.1}$$

donde $\mu(\cdot\ ,\cdot)$ y $v(\cdot)$ son funciones conocidas, $\boldsymbol{\theta}$ y σ^2 son parámetros desconocidos del modelo. Se asume que los y_k, $k \in U$, son condicionalmente independientes dadas las $\mathbf{x}_k$. Nótese que ν puede ser una función conocida de μ como en los modelos lineales generalizados.

Los estimadores óptimos, asistidos por un modelo de super-población ξ, que minimizan el valor esperado de la varianza basada en un diseño de muestreo, $E_\xi(Var_p(\hat{Y}))$, han sido discutidos[8] por muchos autores. Por ejemplo, en Isaki & Fuller (1982*b*) esta varianza esperada tomó el nombre de varianza anticipada.

Resultado 10.9.1 (Teorema 1 de Wu (2003))**.** *Sea t_{y,C_u} un estimador de calibración del total poblacional de la característica de interés, construido utilizado la restricción (10.9), donde $C_u = \{u(\mathbf{x}_1), u(\mathbf{x}_2), \ldots, u(\mathbf{x}_N)\}$ es la familia de vectores de todas las posibles funciones de valor real aplicadas a la información auxiliar.*

[6] Nótese que si el vector de información auxiliar tiene P variables auxiliares, entonces habrán P ecuaciones de calibración.

[7] Bajo este marco de referencia aparece una reducción en la cantidad de restricciones que se utilizan en la calibración.

[8] Los términos E_p y Var_p se refieren a la esperanza y varianza bajo un diseño muestral $p(\cdot)$, y E_ξ y Var_ξ denotan la esperanza y varianza bajo un modelo de super-población ξ.

Dentro de la clase de estimadores de calibración t_{y,C_u}, la escogencia de

$$C_\mu = \{\mu(\mathbf{x}_1, \boldsymbol{\theta}), \mu(\mathbf{x}_2, \boldsymbol{\theta}), \ldots, \mu(\mathbf{x}_N, \boldsymbol{\theta})\}$$

minimiza $E_\xi(Var_p(\hat{Y}))$ bajo el modelo de super-población dado por (10.9.1) y suponiendo condiciones de regularidad en el diseño de muestreo.

Con este resultado podemos proseguir a la construcción del estimador óptimo de calibración resultante de minimizar Ji-cuadrado sujeta a la siguiente restricción

$$\sum_{k\in S} w_k \hat{\mu}_k = \sum_{k\in U} \hat{\mu}_k$$

Donde $\hat{\mu}_k = \mu(\mathbf{x}_k, \hat{\boldsymbol{\theta}})$. La razón para esto se debe a que los valores del vector $\boldsymbol{\theta}$ son desconocidos y se deben reemplazar por un estimador basado en la muestra seleccionada dado por $\hat{\boldsymbol{\theta}}$. La minimización se realiza usando un multiplicador de Lagrange como en Deville (1999). De esta manera, es muy fácil conseguir la expresión del estimador óptimo de calibración, el cual está dado por (Wu & Sitter 2001*b*)

$$\begin{aligned}\hat{t}_{y,opt} &= \sum_{k\in S} w_k y_k \\ &= \hat{t}_{y\pi} + (t_{\hat{\mu}} - \hat{t}_{\hat{\mu}\pi})\hat{B}_y\end{aligned}$$

en donde $t_{\hat{\mu}} = \sum_{k\in U} \hat{\mu}_k$ es el total poblacional de las funciones $\hat{\mu}$, $\hat{t}_{\hat{\mu}\pi}$ su correspondiente estimador de Horvitz-Thompson y

$$\hat{B}_y = \frac{\sum_{k\in S} d_k q_k \hat{\mu}_k y_k}{\sum_{k\in S} d_k q_k \hat{\mu}_k^2}$$

En resumen, los estimadores óptimos de calibración se han estudiado y profundizado en Wu & Sitter (2001*b*) y Wu (2003) y su fundamento se encuentra en la inferencia asistida por modelos. Para motivar las condiciones de optimalidad se utilizó un modelo de super-población semi-paramétrica general dado por (10.9.1). Estos estimadores de calibración para el total poblacional de la característica de interés tiene las siguientes características:

1. Una distancia Ji-cuadrado cuyos factores de peso satisfacen $q_k > 0$ y además sean tales que $N^{-1}\sum_{k=1}^N q_k^2 = O(1)$.
2. Una sola restricción, dada por una reducción de dimensión $u_k = \mu(\mathbf{x}_k, \boldsymbol{\theta})$, donde la forma funcional $\mu(\cdot\ ,\cdot)$ puede ser arbitraria.

Algunos de los resultados más importantes de este método pueden ser resumidos de la siguiente manera (Wu 2003):

- Sea $\hat{\boldsymbol{\theta}} = (\sum_{k\in S} d_k q_k \mathbf{x}_k \mathbf{x}_k')^{-1} \sum_{k\in S} d_k q_k \mathbf{x}_k y_k$. Si se usa $u_k = \mathbf{x}_k'\boldsymbol{\theta}$ como variable de calibración, el estimador de calibración resultante es idéntico al estimador convencional de calibración dado por $\hat{t}_{ycal}$. Por tanto la clase de estimadores resultantes de este método es muy general pues incluye al estimador original como un caso particular.

- Para cualquier estimador consistente de $\boldsymbol{\theta}$ tal que $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta} + o_p(1)$, si se reemplaza $\boldsymbol{\theta}$ por $\hat{\boldsymbol{\theta}}$, en las ecuaciones de calibración, el estimador de calibración resultante no cambia asintóticamente.
- Los estimadores óptimos de calibración obtenidos usando $u_k = E_\xi(y_k \mid \mathbf{x}_k) = \mu(\mathbf{x}_i, \boldsymbol{\theta})$ son óptimos bajo el criterio del mínima varianza esperada.
- Los estimadores óptimos de calibración son óptimos bajo el modelo de superpoblación ξ, pero aun si el modelo considerado es incorrectamente especificado, estos estimadores permanecen consistentes.

Dado que no existe un estimador insesgado con varianza mínima uniforme, la única escogencia de $u(\cdot)$ que hace a $\hat{t}_{yopt}$ un estimador con las anteriores características es $u(\mathbf{x}_k) = y_i$, y por supuesto esto es prácticamente inútil. Por tanto se debe hacer $u(\mathbf{x}_k) \approx y_k$.

El lector debe notar que la estructura del modelo ξ dado por (10.9.1) es muy general e incluye dos importantes casos: el primero el modelo de regresión lineal o no lineal dado por

$$y_k = \mu(\mathbf{x}_k, \boldsymbol{\theta}) + \nu_k \varepsilon_k \tag{10.9.2}$$

donde los ε_k son variables aleatorias independientes e idénticamente distribuidas con $E_\xi(\varepsilon_k) = 0$, $Var_\xi(\varepsilon_k) = \sigma^2$ y $\nu_k = \nu(\mathbf{x}_k)$ es una función conocida y estrictamente positiva.

El segundo caso se refiere al modelo lineal generalizado dado por

$$g(\mu_i) = \mathbf{x}_k'\boldsymbol{\theta}, \quad Var_\xi(y_k|\mathbf{x}_k) = \nu(\mu_k) \tag{10.9.3}$$

donde $\mu_k = E_\xi(y_k|\mathbf{x}_k)$, $g(\cdot)$ es una función de vínculo y $\nu(\cdot)$ es una función de varianza.

A continuación se describe el comportamiento de los estimadores óptimos de calibración bajo un modelo lineal y un modelo log-lineal.

$u(\mathbf{x})$ Vía mínimos cuadrados

Si la información auxiliar explica a la característica de interés de forma lineal, como se observa en la figura 10.5, entonces tendría sentido el argumento que se expresa en Deville & Särndal (1992), en donde motivados por el estimador de razón, se argumenta que ≪*...las ponderaciones [de calibración] que se ajustan bien a las variables auxiliares [reproducen exactamente su total poblacional], también se ajustan bien a la variable de estudio...*≫

En el caso multivariado, la función que hace óptimo al estimador de calibración está dada por

$$u(\mathbf{x}_k, \boldsymbol{\theta}) = \mathbf{x}_k'\boldsymbol{\theta} = \theta_0 + \theta_1 x_{k1} + ... + \theta_P x_{kP} \tag{10.9.4}$$

en donde $\boldsymbol{\theta} = (\theta_0, \theta_1, ..., \theta_P)$ es estimado a través de mínimos cuadrados ponderados, como en una regresión múltiple. Por lo tanto la característica de interés sigue el siguiente modelo de super-población

$$y_k = \mathbf{x}_k'\boldsymbol{\theta} + \nu_k \varepsilon_k \tag{10.9.5}$$

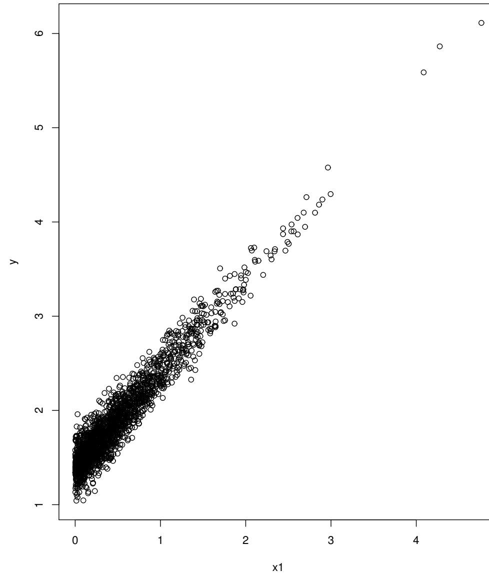


Figura 10.5: *Comportamiento lineal de la característica de interés explicada por la información auxiliar.*

donde los ε_k son independientes e idénticamente distribuidos con $E_\xi(\varepsilon_k) = 0$ y $Var_\xi(\varepsilon_k) = \sigma^2$, y $\nu_k = \nu(\mathbf{x}_k) = 1$. Por tanto al estimar $\boldsymbol{\theta}$ usando la técnica de mínimos cuadrados se tiene que

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= \left(\sum_{k\in S} q_k d_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \sum_{k\in S} q_k d_k \mathbf{x}_k y_k \\ &= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}\end{aligned}$$

donde $\mathbf{V} = diag(d_1 q_1, \ldots, d_n q_n) = \frac{1}{\sigma^2} diag(d_1, \ldots, d_n)$.

Resultado 10.9.2. *De esta forma, el estimador de calibración del total poblacional resultante del anterior modelo de super-población está dado por*

$$\hat{t}_{y,opt} = t_{y\pi} + (\mathbf{t_x} - \hat{\mathbf{t}}_{\mathbf{x}\pi})'\hat{\boldsymbol{\theta}} \tag{10.9.6}$$

Prueba.

$$
\begin{aligned}
\hat{t}_{y,opt} &= \hat{t}_{y\pi} + (t_{\hat{\mu}} - \hat{t}_{\hat{\mu}\pi})\hat{B}_y \\
&= \hat{t}_{y\pi} + (\sum_{k\in U} \hat{\mu}_k - \sum_{k\in U} d_k\hat{\mu}_k)\hat{B}_y \\
&= \hat{t}_{y\pi} + (\sum_{k\in U} \mathbf{x}_k'\hat{\boldsymbol{\theta}} - \sum_{k\in U} d_k\mathbf{x}_k'\hat{\boldsymbol{\theta}})\hat{B}_y \\
&= \hat{t}_{y\pi} + (\sum_{k\in U} \mathbf{x}_k' - \sum_{k\in U} d_k\mathbf{x}_k')\hat{\boldsymbol{\theta}}\hat{B}_y \\
&= \hat{t}_{y\pi} + (\sum_{k\in U} \mathbf{x}_k - \sum_{k\in U} d_k\mathbf{x}_k)'\hat{\boldsymbol{\theta}}\hat{B}_y \\
&= \hat{t}_{y\pi} + (\sum_{k\in U} \mathbf{x}_k - \sum_{k\in U} d_k\mathbf{x}_k)'\hat{\boldsymbol{\theta}}
\end{aligned}
$$

puesto que $\hat{B}_y = 1$. Lo anterior se tiene de la definición de $\hat{B}_y$ teniendo en cuenta que

$$\hat{\mu}_k = \mathbf{x}_k'\boldsymbol{\theta} = x_k'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$$

Por tanto,

$$
\begin{aligned}
\sum_{k\in S} d_kq_k\hat{\mu}_k^2 &= \mathbf{y}'\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \\
&= \mathbf{y}'\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \\
&= \sum_{k\in S} d_kq_k\hat{\mu}_ky_k
\end{aligned}
$$

■

Nótese que el termino $\hat{B}_Y$ es igual a uno y por tanto desaparece, lo que hace que el estimador óptimo de calibración sea idéntico al estimador de calibración clásico dado por (10.4.5).

$u(\mathbf{x})$ Vía modelo lineal generalizado

¿Qué sucede si la información auxiliar no describe a la característica de interés con un comportamiento lineal?, como se observa en la figura 10.6

Es ésta la parte más importante del desarrollo práctico en los estimadores óptimos de calibración. Al respecto, el usuario puede pensar por un instante en los siguientes cuestionamientos:

- Si una característica de información auxiliar explica muy bien a la característica de interés, entonces calibrar con respecto a esta información auxiliar sería muy conveniente. Sin embargo, esta relación no siempre será lineal.
- Si queremos estimaciones perfectas deberíamos utilizar a la misma característica de interés para calibrar, pero como esto es un absurdo se debe utilizar $u(\mathbf{x})$ semejante a y.

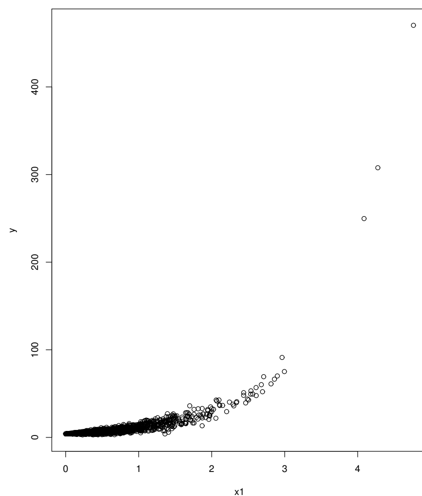


Figura 10.6: *Comportamiento no lineal de la característica de interés explicada por la información auxiliar.*

Si se conoce que la información auxiliar disponible no describe a la característica de interés de forma lineal, se ponen en tela de juicio la aplicación de los estimadores clásicos de calibración motivadas por Deville (1999). Por tanto, si los valores de la característica de interés son considerados como realizaciones de un modelo de super-población ξ como en (10.9.1) que puede ser descrito a través de su primer y segundo momento, entonces claramente el modelo lineal generalizado (MLG), descrito detalladamente en McCullagh & Nelder (1989) y dado por (10.9.3). La mayor particularidad del MLG es que la varianza de la característica de interés depende de la media μ_k. Además, en el MLG se considera que la característica de interés se relaciona con las variables de información auxiliar mediante la media μ_k y una función de vínculo $g(\cdot)$ tal que

$$g(\mu_k) = \theta_0 + \theta_1 x_{k1} + ... + \theta_P x_{kP}$$

Nótese que el modelo clásico de regresión lineal es un caso particular del MLG en donde $g(\mu_k) = \mu_k$ y $V(\mu_k) = 1$. Por supuesto, existen otras formas de la función de varianza y, vínculos no lineales también son permitidos. Por ejemplo, entre las funciones de vínculo y de varianza más populares están el vínculo logarítmico dado por $g(\mu_k) = \log(\mu_k)$ y las funciones de varianza de Poisson dada por $V(\mu_k) = \mu_k$ y la varianza Gamma dada por $V(\mu_k) = \mu_k^2$.

El MLG es un método semi-paramétrico y requiere especificaciones solamente en el primer y segundo momento. La función de vínculo μ_k está relacionada a las variables independientes y la función de varianza describe cómo la variación en la característica de interés está relacionada con la media.

Los coeficientes $(\theta_0, \theta_1, ..., \theta_k)$ pueden ser estimados, como en nuestro caso, usando el método de máxima cuasi-verosimilitud. Para el caso más general, el estimador del vector de parámetros poblacionales $\hat{\boldsymbol{\theta}} = (\hat{\theta_0}, \hat{\theta_1}, ..., \hat{\theta_P})'$, es la solución de la

siguiente ecuación

$$\mathbf{D}'\mathbf{V}^{-1}(\mathbf{y}-\boldsymbol{\mu})=0 \tag{10.9.7}$$

La anterior, no es más que una generalización de las ecuaciones normales en un modelo de regresión múltiple. Donde $\mathbf{y}=(y_1,...,y_n)'$ y $\boldsymbol{\mu}=(\mu_1,...,\mu_n)'$, $\mathbf{V}=diag(V(\mu_1),...,V(\mu_n))$ son las estructuras de media y varianza del modelo respectivamente, y $\mathbf{D}=\partial\mu/\partial\theta$. Los parámetros θ_p, $p=1,\ldots,P$, se encuentran implícitos en (10.9.7). En el caso más simple, el modelo lineal clásico, se tiene que $\mu_k=\theta_0+\theta_1x_{k1}+...+\theta_Px_{kP}$, $\boldsymbol{\mu}=\mathbf{X}'\boldsymbol{\theta}$ y $\mathbf{D}=\mathbf{X}'$. Luego, (10.9.7) queda convertida en $\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\theta}=\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$, las cuales corresponden a las ecuaciones normales de la regresión múltiple.

Por otro lado, en cualquier otro modelo, en donde la función de vínculo sea distinta de la identidad, la mayor dificultad para encontrar el estimador máximo cuasi-verosímil de $\boldsymbol{\theta}$ es que para resolver (10.9.7) se necesita utilizar procedimientos iterativos.

Resultado 10.9.3. *Bajo un modelo de super-población MLG, el estimador óptimo de calibración está dado por*

$$\hat{t}_{y,opt}=\hat{t}_{y\pi}+(t_{\hat{\mu}}-\hat{t}_{\hat{\mu}\pi})\hat{B}_y \tag{10.9.8}$$

con

$$\hat{B}_y=\frac{\sum_{k\in S}d_kq_k\hat{\mu}_ky_k}{\sum_{k\in S}d_kq_k\hat{\mu}_k^2}$$

donde $\hat{\mu}_k=g^{-1}(\mathbf{x}_k'\boldsymbol{\theta})$ *y* $g^{-1}(\cdot)$ *es la inversa de la función de vínculo.*

El software estadístico R tiene implementada la función `glm`, la cual permite estimar los parámetros del MLG. Suponga que se desea encontrar el estimador de máxima cuasi-verosimilitud de $\boldsymbol{\theta}=(\theta_0,\theta_1,...,\theta_P)'$ para el modelo

$$\mu_k=\exp(\theta_0+\theta_1x_{k1}),\quad Var_\xi(y_k|\mathbf{x}_k)=\nu(\mu_k)^2=\mu_k^2 \tag{10.9.9}$$

Por supuesto, desde (10.9.3), se tiene que la función de vínculo es el logaritmo. Las siguientes líneas de código muestran cómo obtener $\hat{\boldsymbol{\theta}}$

```
theta0 <- lm(Y ~ X)
theta1 <- glm(Y ~ X, start=theta0,quasi(var="mu^2",link="log"))
```

Nótese que `theta0` es el estimador de mínimos cuadrados y sirve como estimador inicial para el proceso iterativo. Análogamente, es posible crear un código propio para computar las estimaciones del vector de parámetros basado en McCullagh & Nelder (1989, p. 327).

```
tol <- 0.000000001
theta0 <- solve(t(X)%*%X,t(X)%*%(Y))  ## valores iniciales
dif <- 1
while(dif>=tol) ## condición de iteración
{
mu <- exp(as.vector(X%*%theta0))
V <- diag(1/mu)
```

```
theta1 <- theta0+solve(t(X)%*%X,t(X)%*%V%*%(Y-mu))
dif <- max(abs(theta1-theta0))
theta0 <- theta1
}
```

Por supuesto, el anterior código debe coincidir con la salida que arroje el procedimiento `glm` de R.

10.10 Ejercicios

10.1 (Tille & Ardilly 2006, Ejercicio 7.1) Usando un procedimiento de muestreo, se obtuvieron las siguientes estimaciones para los tamaños absolutos $\hat{N}_{ij}$ de dos sub-poblaciones de interés:

80	170	150	400
90	80	210	380
10	80	130	220
180	330	490	1000

Por otro lado, mediante fuentes oficiales, los tamaños marginales sí se conocen con certeza. Los verdaderos totales para las filas son $(430, 360, 210)$ y los totales verdaderos para las columnas son $(150, 300, 550)$.

(a) Ajuste la tabla sobre los verdaderos totales marginales de la población usando el algoritmo IPFP.

(b) Ajuste la tabla sobre los verdaderos totales marginales de la población usando el enfoque de calibración con el método de *raking*.

(c) Explique las diferencias o similitudes entre las anteriores estimaciones.

10.2 (Tille & Ardilly 2006, Ejercicio 7.4) Suponga que se obtuvo la siguiente tabla y que los verdaderos totales para las filas son $(84, 37, 444, 464)$ y los totales verdaderos para las columnas son $(49, 859, 11, 10)$.

78	6	0	0	84
32	5	0	0	37
0	0	427	17	444
0	0	432	32	464
110	11	859	49	1029

Como se puede notar, los totales estimados por fila coinciden plenamente con los verdaderos totales. Explique por qué está tabla no se puede ajustar al utilizar el algoritmo IPFP.

10.3 Considere una region agrícola consistente en $N = 2010$ fincas, para la cual se seleccionó una muestra aleatoria simple de fincas de tamaño $n = 100$. Además, se sabe que hay 1580 fincas con menos de 160 hectáreas (post-estrato 1) y 430 fincas con más de 160 hectáreas (post-estrato 2). La característica de interés medida en cada finca incluida en la muestra es el área de cereal

cultivada en cada finca. Si se tuvo una muestra realizada en donde $n_1 = 70$, $n_2 = 30$, $\bar{y}_1 = 19.4$ y $\bar{y}_2 = 51.63$, estime usando la técnica de calibración, la media poblacional del área de cereal cultivada en la región agrícola y reporte el coeficiente de variación estimado.

10.4 Considere un diseño de muestreo de Poisson con probabilidades de inclusión desiguales π_k, $k \in U$. Suponga que se tiene interés en la estimación del total poblacional t_y. Construya un estimador de calibración usando una sola característica de información auxiliar $x_k = 1$ y $q_k = 1$, para todo $k \in U$, usando la siguiente pseudo-distancia (parametrizada por α):

$$G(x) = \begin{cases} \frac{1}{\alpha(\alpha-1)}(x^\alpha + (\alpha - 1) - \alpha x), & \text{si } \alpha \in \mathbb{R} - \{0, 1\} \\ x\ln(x) + 1 - x, & \text{si } \alpha = 1 \\ \ln(1/x) - 1 + x, & \text{si } \alpha = 0 \end{cases}$$

(a) Escriba las ecuaciones de calibración.

(b) Obtenga la función $g(x)$ para los tres casos de α.

(c) Demuestre que la función $F(u)$ es fija e igual a $N/\hat{N}$.

(d) Deduzca los pesos de calibración.

(e) Obtenga el estimador de calibración resultante. ¿Qué forma tiene el estimador resultante?.

10.5 Suponga que la información del ejercicio 8.7. es el resultado de un plan de muestreo Poisson con probabilidad de inclusión $\pi_k = n(x_k/t_x)$. Utilizando los resultados del ejercicio anterior y suponiendo que $x_k = 1$ y $q_k = 1$, para todo $k \in U$, obtenga una estimación de calibración para el total de habitantes en el municipio, el numero de automóviles en el municipio y el número de efectivos militares en el municipio. Obtenga los correspondientes coeficientes de variación estimados.

10.6 Sustente o refute las siguientes afirmaciones

(a) Los estimadores de calibración inducidos por la distancia Ji-cuadrado coinciden plenamente con los estimadores de regresión general.

(b) La cantidad q_k es constante para todos los individuos bajo la distancia de entropía.

(c) Bajo la distancia Ji-cuadrado inversa, al minimizar la distancia con respecto a las restricciones de calibración, siempre se llega a que los pesos w_k son iguales al inverso de la probabilidad de inclusión del k-ésimo elemento.

Capítulo 11

Inferencia basada en modelos poblacionales

> Cualquier estimador usado en inferencia de poblaciones finitas debería estar basado en el mejor modelo que pueda ser construido convenientemente y, cuando fuese posible, en la inferencia basada tanto en modelos predictivos como en la inferencia basada en el diseño de muestreo.
>
> Ken Brewer (2002)

Los anteriores capítulos de este texto se enfocan en la búsqueda de una estrategia representativa de muestreo bajo el paradigma de la inferencia basada en el diseño de muestreo. Cabe resaltar que este enfoque, propuesto en Neyman (1934), es ampliamente utilizado por las agencias gubernamentales en todo el mundo. Sin embargo, ha sido atacado desde hace varias décadas por estadísticos teóricos que muestran su inconformidad con los fundamentos filosóficos de este enfoque. En las últimas décadas se han propuesto otro tipo de enfoques, el más importante de ellos: el enfoque basado en modelos de super-población. Antes de adentrarnos en este tema, es necesario realizar un pequeño repaso a través de la evolución de la inferencia para poblaciones finitas.

11.1 Un poco de historia

Según Rao (2005), el primer personaje interesado en el método representativo (más adelante conocido como teoría de muestreo) fue el estadístico noruego Kiaer (1901) puesto que demostró empíricamente que seleccionando muestras estratificadas se obtienen mejores resultados en los estimativos de medias y totales. Más adelante, Bowley (1906) utiliza aproximaciones a la distribución normal para la estimación de proporciones y propone la fórmula de la estimación de la varianza para diseños de muestreo estratificados. Para la década de 1920, el método representativo era usado de manera difundida en Estados Unidos y alrededor del mundo. Fue así como en 1924, el ISI (Instituto Internacional de Estadística, por su traducción del inglés al español) crea una comisión de discusión de este método. Los resultados de

este comité incluyen el trabajo de Bowley (1926) basado en métodos de selección representativos con probabilidades de inclusión iguales. Con estos avances teóricos y con la publicación de tablas de números aleatorios por Tippett (1927) se facilitó la selección de muestras probabilísticas. En ese mismo año, Hubback (1927)[1] reconoce la necesidad de utilizar este enfoque en los estudios agrícolas puesto que:

1. Evita los posibles sesgos personales.
2. Es posible determinar un tamaño de muestra tal que satisfaga un margen de error determinado por el investigador.

El trabajo de Bowley junto con el reporte del ISI hicieron que Neyman examinara las mismas bases de la inferencia en poblaciones finitas. En particular, el artículo de Neyman (1934) es considerado como uno de los pilares en donde descansan los fundamentos del muestreo como se conoce hoy en día. Al respecto Leslie Kish, en un comentario al artículo de Smith (1976), asegura que Neyman hizo siete grandes contribuciones al muestreo:

1. Propuso la asignación de Neyman para el tamaño de muestra con diseños estratificados.
2. Descubrió que el muestreo por conglomerados puede realizarse basado en un esquema probabilístico tal que las varianzas de los estimadores resultantes pudieran ser calculadas o estimadas.
3. Para que lo anterior se tuviera, se necesita una muestra grande de unidades.
4. Para seleccionar una muestra grande es crucial definir un marco de selección de números aleatorios.
5. El conocimiento subjetivo del comportamiento de la población puede usarse para formar subgrupos poblacionales o estratos.
6. Un esquema de selección probabilístico es mejor que un esquema de selección a conveniencia.
7. Para convencer a los escépticos acerca de la validez de sus afirmaciones, se dispuso a realizar ejemplos prácticos con encuestas verdaderas a gran escala.

La nueva teoría de Neyman revolucionó el mundo del muestreo y lo liberó del paradigma de las probabilidades de inclusión iguales. Él introdujo, en un solo artículo, las ideas de eficiencia, asignación óptima, generalización del teorema de Markov, muestreo por conglomerados y presentó un caso evidente en donde, utilizando el muestreo a conveniencia, se llegaba a conclusiones equivocadas. Más adelante, Neyman propuso el muestreo en dos fases. Smith (1976) afirma que el muestreo con probabilidad proporcional y el muestreo en varias etapas son resultado de las ideas de Neyman. Asimismo, propuso realizar la inferencia para muestras grandes basado en la teoría de los intervalos de confianza ≪*sin tener en cuenta las*

[1]El trabajo inicial de R. A. Fisher fue influenciado por Hubback. Nótese que Fisher justificó el análisis de los datos sin tener en cuenta el método de selección solamente en los casos en que los resultados fueran muy cercanos al considerar la aleatorización de las unidades (Smith 1976).

propiedades de la población finita, fuesen cuales fuesen≫. Cualquier método que satisficiera los anteriores supuestos era llamado representativo.

Cochran (1939) hace varios aportes significativos: introduce el uso del ANOVA para estimar la ganancia en eficiencia debida a la estratificación, propone la estimación de la varianza para encuestas en dos etapas y reúne los componentes para realizar estimación por regresión bajo muestreo en dos fases. También introduce el concepto de super-población: ≪*La población finita podría ser vista como una muestra aleatoria de una población infinita*≫. Más adelante, Cochran (1940) introduce el estimador de razón y desarrolla la teoría de la estimación de totales y medias mediante modelos de regresión. Poco tiempo después Madow & Madow (1944) introducen la teoría del muestreo sistemático.

Mientras tanto en India, Mahalanobis funda el Instituto Indio de Estadística en donde hace grandes aportes formulando expresiones de la varianza de estimadores en función del costo de una encuesta. Varios textos surgieron después de la década de 1940 en donde se trataba el problema de la selección de muestras y la estimación de parámetros en poblaciones finitas. Uno de los mas grandes desarrollos en términos de la teoría actual lo hicieron Horvitz & Thompson (1952) quienes propusieron un marco de trabajo para la teoría del muestreo proporcional sin reemplazo y desarrollaron un elegante tratamiento del muestreo, completando así las bases de la inferencia basada en el diseño de muestreo.

Definición 11.1.1. ***La inferencia basada en el diseño (Design-Based)*** *(Särndal, Swensson & Wretman 1992) Este enfoque realiza las estimaciones de los parámetros dependiendo del diseño de muestreo escogido para seleccionar la muestra sin tener en cuenta las propiedades de la población finita. Así por ejemplo el estimador del total poblacional t_y estará dado por:*

$$\hat{t}_y = \sum_{k \in m} d_k y_k$$

en donde d_k es una ponderación inducida por el diseño de muestreo. Bajo esta perspectiva, los valores y_k son tomados como la observación en el individuo k de la característica de interés y. Sin embargo, y no se toma como una variable aleatoria, sino como una cantidad fija.

Desde ese punto hasta nuestros días han aparecido avances y aportes y nuevas teorías de selección de muestras y de estimación de parámetros manteniendo la filosofía de la inferencia basada en el diseño de muestreo. Rao (2005) cita algunas de ellas por ejemplo: muestreo en varias ocasiones, muestras tipo panel, estimación de funciones de distribución y cuantiles, estimación en dominios pequeños.

Por otro lado, al mismo tiempo Godambe (1955) prueba el siguiente teorema que pone en tela de juicio el concepto de eficiencia al que Neyman se refería puesto que prueba que, bajo la inferencia basada en el diseño de muestreo, no existe un estimador insesgado de varianza mínima.

Resultado 11.1.1. *Sea un diseño muestral $p(\cdot)$ de tamaño de muestra $n(S) < N$ tal que $\pi_k > 0 \ \ \forall k \in U$. Entonces, no existe un estimador insesgado de varianza mínima uniformemente en la clase de todos los estimadores insesgados*

Prueba. Basu (1971) propone la siguiente demostración: suponga que $\hat{t}$ es un estimador insesgado para el total poblacional t. Por tanto, $\hat{t}$ es insesgado para cualquier estructura poblacional $\mathbf{y} = (y_1, \ldots, y_N)$. Nótese que $\mathbf{y} \in \mathcal{Y}$, con $\mathcal{Y}$ el conjunto de todas las posibles poblaciones. En particular, este estimador es insesgado para $\mathbf{y}_0 \in \mathcal{Y}$. Por tanto $\hat{t_0}$ es insesgado para t_0. Ahora, nótese que

$$\hat{t}^* = \hat{t} + t_0 - \hat{t}_0 \tag{11.1.1}$$

es también un estimador insesgado para t. Además cuando $\mathbf{y} = \mathbf{y}_0$, se tiene que $\hat{t}^* = t_0$ y por tanto $Var(\hat{t}^*) = 0$. En conclusión, para que un estimador insesgado sea de mínima varianza uniformemente para cualquier estructura poblacional $\mathbf{y} \in \mathcal{Y}$ debe tener varianza nula. Lo anterior es imposible puesto que el diseño de muestreo no considera el censo. Por tanto, se tiene la demostración del resultado. ■

El anterior teorema es resultado de la generalidad del enfoque inferencial de Neyman puesto que las inferencias son realizadas con respecto al diseño de muestreo sin tener en cuenta la estructura poblacional de la población. Smith (1976) afirma que este enfoque permite mucha libertad para una teoría inferencial y por ende no es posible encontrar un óptimo para todas las estructuras poblacionales. El anterior argumento sumado a la fábula de los elefantes de Basú (Basu 1971) hicieron que los estadísticos teóricos se replantearan el seguir haciendo inferencias basadas en el diseño de muestreo.

La fábula de los elefantes de Basú

Como lo afirma Brewer (2002), la siguiente fábula publicada removió las bases de la inferencia basada en el diseño de muestreo.

> El dueño de un circo está planeando transportar sus 50 elefantes adultos, para este propósito él necesita una buena estimación del peso total de los elefantes. Como pesar un elefante es una tarea muy incómoda, el dueño del circo quiere estimar el peso total pesando sólo un elefante. ¿Cuál elefante debería pesar? El dueño del circo decide echar un vistazo a sus registros y descubre una antigua lista de los pesos de los elefantes, elaborada hace tres años. Él encuentra que tres años atrás Sambo, un elefante mediano, era el promedio (en peso) de su manada. El dueño del circo verifica la información con el entrenador quien le asegura que Sambo todavía puede considerarse como el promedio de la manada.
>
> Así, el dueño del circo planea pesar a Sambo y tomar a $50 \times y_{Sambo}$ (donde y_{Sambo} es el peso de Sambo) como una estimación del peso total $t_y = y_1 + \ldots + y_{50}$ de la manada. Pero el estadístico del circo se aterra al conocer la estrategia de muestreo del dueño (con un diseño de muestreo no probabilístico).
>
> - ¿Cómo puede obtener una estimación insesgada de t_y? - protesta el estadístico.
>
> Así, ellos trabajan juntos en la elaboración de un diseño de muestreo. Con la ayuda de una tabla de números aleatorios, construyen un

> plan que asigna una probabilidad de inclusión de 99/100 a Sambo y probabilidades de 1/4900 al resto de la manada. Naturalmente, Sambo es seleccionado y el dueño del circo está feliz.
>
> - ¿Cómo va a estimar t_y? - pregunta el estadístico - ¿Por qué? La estimación debería ser $50 \times y_{Sambo}$, por supuesto, responde el dueño. - ¡Oh¡ no, eso es incorrecto, responde el estadístico. recientemente, yo leí en un artículo de Annals of Mathematical Statistics, en donde se prueba que el estimador de Horvitz-Thompson es el *único estimador hiper-admisible en la clase de todos los estimadores insesgados polinomiales generalizados.* - ¿Cuál sería la estimación de Horvitz-Thompson en este caso?, pregunta el impresionado dueño - Dado que la probabilidad de inclusión de Sambo fue de 99/100, dice el estadístico, el estimador es $\dfrac{100}{99} \times y_{Sambo}$.
>
> - ¿Y cuál sería nuestra estimación si el plan de muestreo hubiese seleccionado a Jumbo?, pregunta el incrédulo dueño. - De acuerdo a lo que yo entiendo acerca del estimador de Horvitz-Thompson, dice el infeliz estadístico, el estimador de t_y sería $4900 \times y_{Jumbo}$ donde y_{Jumbo} es el peso de Jumbo.
>
> De esta forma, el estadístico perdió su empleo (y quizás se convirtió en profesor).

Lohr (2000) se pregunta si fue justo despedir o no al estadístico. Una de las fallas en la estrategia de muestreo utilizada por el estadístico se presenta en la construcción del diseño de muestreo que induce una probabilidad de selección muy grande a un elefante cuyo valor de la característica de interés, el peso, es promedio con respecto al resto de la manada. Como se vio en capítulos anteriores, una estrategia de muestreo eficiente, que utilice al estimador de Horvitz-Thompson, es aquella cuyo diseño de muestreo induzca probabilidades de inclusión proporcionales al valor que toma la característica de interés.

La fábula de los dos estadísticos

Lahiri (1968) expresa las dificultades que sobrevienen al tratar de explicarle a un hombre del común el enfoque de la inferencia en poblaciones finitas mediante la siguiente situación

> Suponga que dos estadísticos (¿muestristas?) son contratados para seleccionar una muestra de tamaño n de una determinada población finita. Ambos poseen la misma información acerca del comportamiento de la población. Este conocimiento incluye una característica de información auxiliar para cada unidad perteneciente a la población. Uno decide seleccionar una muestra aleatoria simple y el otro decide seleccionar una muestra con probabilidad proporcional al tamaño. Como complemento de la estrategia de muestreo, ambos deciden utilizar el estimador $\bar{y} = \sum_s y_k/n$. De manera increíble, los dos estadísticos seleccionan exactamente las mismas unidades en la muestra de tamaño n. Por supuesto, ambos saben que la desviación típica está dada en términos de $\bar{y} - \bar{Y}$; sin embargo, ambos proponen medidas totalmente diferentes para la precisión de sus estimadores.

¿Cómo explicar esta situación? Este tipo de fábulas hacen un gran aporte al desarrollo de la estadística. De hecho, la narración anterior es un claro ejemplo de como en nuestra ciencia estadística hay una gran cantidad de cosas por hacer. Sin embargo, nótese que el mismo tipo de razonamiento aparece si los mismos estadísticos anteriores se enfrentaran a un problema frecuentista y uno de ellos decide que la verosimilitud de los datos es normal y el otro decide que es una beta. Seguramente llegarían a distintas estimaciones. Quien propone la estrategia de muestreo se ve obligado a tomar las mismas decisiones subjetivas de quien propone una verosimilitud, en el caso frecuentista, o una distribución a priori, en el caso bayesiano. Ahora, es deber de investigador asegurarse que la subjetividad esté enmarcada dentro de ciertos límites. Por supuesto, si usted va a medir la distancia de la tierra a la luna, seguramente no utilizaría un metro.

Con los anteriores argumentos nació otro tipo de inferencia para poblaciones finitas: el enfoque basado en un modelo de super-población que asume que la estructura de la población obedece a un modelo específico. La distribución inducida por el modelo provee las herramientas para predecir valores particulares específicos para los individuos que no fueron seleccionados en la población.

Definición 11.1.2. ***La inferencia basada en el modelo (model based)*** *(Valliant, Dorfman & Royall 2000, Smith 1976) Este enfoque supone uso de información auxiliar y que relaciona a la característica de interés con la información auxiliar mediante un modelo de superpoblación ξ. Bajo esta perspectiva no se requiere que los datos provengan de una muestra probabilística (la forma en la que se escoge la muestra no se tiene en cuenta para la estimación de los parámetros de interés) y la observación de la característica de interés en las unidades poblacionales y_k se define como la realización de una variable aleatoria Y_k. Partiendo de que el total poblacional se puede escribir como*

$$T_y = \sum_{k \in s} Y_k + \sum_{k \notin s} Y_k, \tag{11.1.2}$$

la tarea es estimar por medio del modelo ξ, las respectivas observaciones y_k de los elementos que no fueron seleccionados en la muestra. Denotando esta estimación como $E(Y_k)$, un predictor para el total estaría dado por:

$$\hat{T}_y = \sum_{k \in s} Y_k + \sum_{k \notin s} E_\xi(Y_k) \tag{11.1.3}$$

y por tanto la realización de $\hat{T}_y$ con los datos específicos de la muestra seleccionada s estaría definida como

$$\hat{t}_y = \sum_{k \in s} y_k + \sum_{k \notin s} \hat{E}_\xi(Y_k) \tag{11.1.4}$$

donde $\hat{E}_\xi(Y_k)$ es una estimación de $E_\xi(Y_k)$ realizada con los datos obtenidos de la muestra seleccionada s.

Godambe & Thompson (1977) sugirieron, en el curso de una discusión en el congreso internacional de estadística en Nueva Delhi, que se debía buscar una manera

de encontrar estimadores que tuvieran sentido en ambos tipos de inferencia. Más adelante, Sarndal & Wright (1984) y Brewer (1999) llevaron a cabo la implementación de esta sugerencia.

Aunque el tipo de inferencia dominante después de la segunda guerra mundial fue la inferencia basada en el diseño de muestreo, a comienzos de los años setenta, Richard Royall, con la ayuda de muchos co-autores, cambiaron rotundamente esa tendencia con gran determinación. Él afirmó que la inferencia basada en el diseño, aunque no hace supuestos acerca de las probabilidades y parece ser no paramétrica y robusta, estaba sujeta a importantes defectos. Algunas de las limitaciones que cita Royall (1971) son:

- Las sorprendentes complicaciones encontradas en el estudio y ejecución de los diseños de probabilidad proporcional al tamaño y
- las torpezas y equivocaciones de casi todos las estimaciones probabilísticas concernientes a la estimación de razones.

La sugerencia de Royall fue aún más radical. Él propuso abandonar la inferencia basada en el diseño de muestreo a favor de estimadores cuyas útiles propiedades (insesgamiento, consistencia, optimalidad, etc.) estuvieran definidas en términos del modelo predictivo apropiado. Esto significa que conceptos como el sesgo y la varianza ya no están definidos como esperanzas a través de todas las posibles muestras, sino como promedios de las realizaciones de las unidades poblacionales (estén en la muestra o no) bajo el modelo predictivo establecido. Desde el punto de vista de Royall, el proceso de aleatorización se convierte en irrelevante y propone que la muestra sea escogida a conveniencia (lo que en la práctica significa escoger las unidades más grandes). Sin embargo, este tipo de inferencia debe ser usado con mucho cuidado pues, como lo afirma Box (1979):

> Todos los modelos son errados, pero algunos son útiles. El hecho de que todos los modelos están equivocados se hace más y más claro cuando el tamaño de la muestra se incrementa; por eso las estimaciones resultantes de un modelo predictivo errado son deficientes.

De una cosa hay que estar seguros la inferencia basada en modelos predictivos y la inferencia basada en el diseño de muestreo no se deben ver como competencia sino como puntos de vista que pueden llegar a ser complementarios, es así como nace la inferencia basada en el diseño de muestreo, pero asistida por modelos predictivos (model assisted survey sampling, en el inglés original). Sin embargo, estos dos tipos de inferencia, aunque se pueden combinar, no se pueden conciliar porque su filosofía es literalmente distinta.

La inferencia basada en el diseño de muestreo difiere radicalmente de la inferencia basada en los modelos predictivos y quizás de cualquier otro modelo estadístico, porque está basada exclusivamente en las observaciones muestrales y no hace supuestos a priori, además su dirección de análisis va en contravía con la dirección de la inferencia basada en modelos. Kyburg (1987) escribe en su artículo una defensa y vindicación sobre la inferencia basada en modelos y hace un comentario con respecto al tipo de inferencias estadísticas que existen; él afirma que:

> La inferencia inversa procede de lo particular a lo general, la inferencia directa de lo general a lo particular.

Desde este punto de vista, la inferencia basada en el diseño de muestreo es inversa y la inferencia basada en modelos predictivos es directa. Nótese que la inferencia bayesiana también pertenece al grupo de las inferencias inversas. Brewer (1999) argumenta que:

> En esta época, la tendencia es usar la inferencia basada en el diseño para la estimación en grandes dominios y muestreo sintético (inferencia basada en modelos) para la estimación en dominios pequeños dentro del mismo estudio.

También hace alusión al uso de los estimadores de calibración cosméticos que combinan los dos tipos de inferencias simultáneamente. La idea de los estimadores cosméticos nace con Sarndal & Wright (1984) y el argumento para utilizar esa palabra es el hecho de que un estimador pueda ser visto o interpretado como un predictor obtenido de una regresión lo hace muy atractivo.

Finalmente, desde la aparición del clásico libro de muestreo de Särndal, Swensson & Wretman (1992), la historia de la inferencia en poblaciones finitas ha tomado otro matiz, definiendo no solamente al blanco y el negro sino que también una especie arco iris entre estas dos corrientes del pensamiento inferencial. Isaki & Fuller (1982*a*) plantean el problema de tener en cuenta la forma de selección de la muestra y a la vez el modelo de relación ξ entre la característica de interés y la información auxiliar, pero es en Cassel, Särndal & Wretman (1976*b*) que se acuña un término muy polémico, la **inferencia asistida por un modelo y basada en el diseño** (model assisted design based, en el inglés original). Es decir, la base de la inferencia es el diseño de muestreo, pero la estrategia de muestreo es complementada teniendo en cuenta un ξ modelo en la estimación del parámetro de interés.

Para terminar el repaso por la historia, en Brewer (2002) se presenta el siguiente diálogo entre dos estadísticos, llamados E y L, que utilizan enfoque de inferencia en poblaciones finitas distintos. Uno utiliza la inferencia basada en modelos predictivos, que utiliza los datos de la muestra para construir un modelo que permita predecir los valores no observados en la muestra y así llegar a una estimación de las cantidades de interés y que no utiliza las probabilidades de inclusión. Y otro que utiliza la inferencia basada en el diseño de muestreo. Cada uno acérrimo defensor de su punto de vista.

> E: Creo que usted sigue viviendo en los ochenta. No tenga la menor duda de que las cosas han cambiado un poco. Muchos estadísticos académicos están a favor de la inferencia basada en modelos predictivos.
>
> L: Es cierto, pero ese tipo de inferencias no es usado en la vida práctica profesional. ¡Dígame al menos una entidad estatal que la utilice!
>
> E: Claro que las hay, al tratar de estimar parámetros en dominios pequeños se utilizan estimadores sintéticos. Esos estimadores están basados en modelos predictivos.

L: Ah, pero sólo se utilizan en dominios pequeños. De lo contrario no se utilizan. Bien, si usted está tratando de estimar un parámetro en un dominio pequeño, la inferencia basada en modelos predictivos puede ser particularmente útil.

E: No, es más que eso, se trata de que la inferencia basada en el diseño muestral es particularmente mala para muestras pequeñas. Fíjese que con una muestra probabilística usted puede seleccionar las unidades más grandes y dejar a las chicas de lado, con la inferencia basada en el diseño tendría unas malas estimaciones. Una manera más segura de evadir esa posibilidad es dividir la población en grupos y hacer una selección de unidades en cada grupo.

L: ¿Cómo una clase de estratificación?

E: Hmm estratificación sí, digamos que sí. La estratificación por el tamaño de las unidades es muy útil, pero el punto es que debería conocer muy bien la población.

L: Precisamente, y si usted no conoce muy bien la población, podría ajustar un modelo totalmente equivocado y como resultado tendría unas muy malas predicciones.

11.2 Algunos modelos predictivos

Valliant, Dorfman & Royall (2000) argumenta que no existe ninguna razón de peso para que los principios de la inferencia en poblaciones finitas estén tan alejados del resto de la teoría estadística. De esta forma, el enfoque de inferencia basado en el diseño de muestreo afirma que la aleatorización de las unidades a la muestra es el único principio válido para realizar inferencias en al población finita. Sin embargo, esta rigidez hace que el estadístico se quede sin bases estadísticas para hacer inferencias si los datos no provienen de algún diseño de muestreo. Por supuesto que es válido pensar que el estadístico tiene muchas herramientas que le permiten hacer inferencia sin importar la naturaleza de los datos. Una de esas herramientas es el principio de verosimilitud que afirma lo siguiente (Gelman, Carlin, Stern & Rubin 2004):

> Al momento de realizar inferencias o tomar decisiones sobre un parámetro θ después de que los datos han sido observados, toda la información relevante se encuentra contenida en la función de verosimilitud para los datos observados.

No es difícil constatar que la función de verosimilitud para cualquier diseño de muestreo es la misma y está dada por una función indicadora. Así que, la conclusión de Valliant, Dorfman & Royall (2000) es que, aunque la aleatorización es deseable, no es ni necesaria ni suficiente para realizar inferencia estadística rigurosa. La validez de la inferencia estadística sigue estando en pie con o sin aleatorización. Las siguientes secciones dan cuenta de alguno de los muchos modelos predictivos que se utilizan para situaciones específicas en la inferencia de poblaciones finitas.

11.2.1 Un modelo para el muestreo aleatorio simple

Suponga que $Y_1, \ldots, Y_N$ es una población de variables aleatorias independientes en idénticamente distribuidas. El mecanismo probabilístico que rige a la población está dado por un modelo de superpoblación ξ definido como

$$Y_k = \beta + \varepsilon_k \tag{11.2.1}$$

Donde cada uno de los ε_k $k \in U$ son variables aleatorias independientes e idénticamente distribuidas con media cero y varianza constante σ^2, tales que:

$$\begin{aligned} E_\xi(Y_k) &= \beta \\ Var_\xi(Y_k) &= \sigma^2. \end{aligned} \tag{11.2.2}$$

De esta población, se selecciona una muestra s de tamaño n. De esta forma se tienen los siguientes resultados.

Resultado 11.2.1. *Bajo el modelo 11.2.1, el mejor estimador lineal insesgado de β está dado por*

$$\hat{\beta} = \bar{Y}_s = \frac{1}{n}\sum_{k\in s} Y_k \tag{11.2.3}$$

Prueba. El estimador de β está dado por la minimización de la siguiente función de dispersión

$$D = \sum_{k\in s} \frac{(y_k - \beta)^2}{\sigma^2}.$$

Luego de derivar e igualar a cero, se encuentra fácilmente que $\hat{\beta} = \bar{Y}_s$. Por otro lado, se tiene que

$$E_\xi(\hat{\beta}) = \frac{1}{n}\sum_{k\in s} E_\xi(Y_k) = \beta$$

Utilizando el teorema de Gauss-Markov (Ravishanker & Dey 2002, Resultado 4.4.1) se tiene que $\hat{\beta}$ es el mejor estimador puesto que tiene varianza mínima. ■

Resultado 11.2.2. *Bajo el modelo 11.2.1, el mejor predictor lineal insesgado de T_y y su error cuadrático medio*[2] *(ECM_ξ) están dados por*

$$\hat{T}_y = \frac{N}{n}\sum_{k\in s} Y_k \tag{11.2.4}$$

$$ECM_\xi(\hat{T}_y) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right)\sigma^2 \tag{11.2.5}$$

respectivamente.

[2] Como T_y y $\hat{T}_y$ son variables aleatorias se utiliza el ECM como medida de variabilidad.

Prueba. En primer lugar nótese que

$$\hat{T}_y = \sum_{k \in s} Y_k + \sum_{k \notin s} \hat{\beta} = \sum_{k \in s} Y_k + (N-n)\bar{Y}_s = \frac{N}{n} \sum_{k \in s} Y_k$$

$\hat{T}_y$ es insesgado puesto que

$$E_\xi(\hat{T}_y - T_y) = E_\xi \left(\frac{N}{n} \sum_{k \in s} Y_k - \sum_{k \in U} Y_k \right) = \beta - \beta = 0$$

Por último,

$$\begin{aligned}
ECM_\xi(\hat{T}_y) &= E_\xi \left(\hat{T}_y - T_y \right)^2 \\
&= E_\xi \left(\left[\frac{N}{n} - 1 \right] \sum_{k \in s} Y_k - \sum_{k \notin s} Y_k \right)^2 \\
&= \left[\frac{N}{n} - 1 \right]^2 E_\xi \left(\sum_{k \in s} Y_k \right)^2 - 2 \left[\frac{N}{n} - 1 \right] E_\xi \left(\sum_{k \in s} Y_k \right) E_\xi \left(\sum_{k \notin s} Y_k \right) \\
&\qquad + E_\xi \left(\sum_{k \notin s} Y_k \right)^2 \\
&= \left[\frac{N}{n} - 1 \right]^2 E_\xi \left(\sum_{k \in s} Y_k \right)^2 - 2(N-n)^2\beta^2 + E_\xi \left(\sum_{k \notin s} Y_k \right)^2 \\
&= \left[\frac{N}{n} - 1 \right]^2 E_\xi \left(\sum_{k \in s} Y_k - n\beta \right)^2 + E_\xi \left(\sum_{k \notin s} Y_k - (N-n)\beta \right)^2
\end{aligned}$$

Dado que $E_\xi \left(\sum_{k \in s} Y_k - n\beta \right)^2 = Var_\xi \left(\sum_{k \in s} Y_k \right)$, entonces se tiene que

$$\begin{aligned}
ECM_\xi(\hat{T}_y) &= \left[\frac{N}{n} - 1 \right]^2 Var_\xi \left(\sum_{k \in s} Y_k \right) + Var_\xi \left(\sum_{k \notin s} Y_k \right) \\
&= \left[\frac{N}{n} - 1 \right]^2 n\sigma^2 + (N-n)\sigma^2 \\
&= \frac{N^2}{n} \left(1 - \frac{n}{N} \right) \sigma^2
\end{aligned}$$

■

Nótese que para estimar σ^2 es posible utilizar a S^2. De esta manera, los dos enfoques de inferencia parecen coincidir puesto que la expresiones para el estimador, y su varianza estimada son idénticas, aunque el trasfondo y la interpretación sean distintas. Además, Lohr (2000) afirma que los intervalos de confianza construidos a partir de los dos enfoques también coinciden aunque su interpretación no[3]

[3]Con el enfoque inferencial basado en el diseño de muestreo la interpretación es como sigue:

11.2.2 Un modelo para el muestreo aleatorio estratificado

Suponga que $Y_1, \ldots, Y_N$ es una población de variables aleatorias cuyo comportamiento es distinto en H grupos poblacionales cada uno de tamaño N_h ($h = 1, \ldots, H$) que definen una población estratificada $U = \{U_1, \ldots, U_H\}$. Claramente el tamaño de la poblacional general es $N = N_1 + \cdots, N_H$. El mecanismo probabilístico que rige a la población está dado por un modelo de superpoblación ξ definido como

$$Y_{hk} = \beta_h + \varepsilon_{hk} \tag{11.2.6}$$

Donde el subscrito hk hace referencia a las cantidades asociadas con el k-ésimo elemento dentro del h-ésimo estrato. Cada uno de los ε_{hk} son variables aleatorias independientes e idénticamente distribuidas con media cero y varianza constante σ_h^2 dentro del estrato h, no correlacionados entre estratos, tales que:

$$\begin{aligned} E_\xi(Y_{hk}) &= \beta_h \\ Var_\xi(Y_{hk}) &= \sigma_h^2 \\ Cov_\xi(Y_{hk}, Y_{gl}) &= 0 \quad \text{si } h \neq g. \end{aligned} \tag{11.2.7}$$

De cada uno de los estratos se extrae una muestra s_h de tamaño n_h ($h = 1, \ldots, H$). El tamaño de la muestra general es $n = n_1 + \cdots, n_H$.

Resultado 11.2.3. *Bajo el modelo 11.2.6, el mejor estimador lineal insesgado de β_h ($h = 1, \ldots, H$) está dado por*

$$\hat{\beta}_h = \bar{Y}_{s_h} = \frac{1}{n_h} \sum_{k \in s_h} Y_{hk} \tag{11.2.8}$$

Prueba. El estimador de β_h está dado por la minimización de la siguiente función de dispersión

$$D = \sum_{k \in s_h} \frac{(y_{hk} - \beta_h)^2}{\sigma_h^2}.$$

Luego de derivar e igualar a cero, se encuentra fácilmente que $\hat{\beta}_h = \bar{Y}_{s_h}$. Por otro lado, se tiene que

$$E_\xi(\hat{\beta}_h) = \frac{1}{n_h} \sum_{k \in s_h} E_\xi(Y_{hk}) = \beta_h$$

Del análisis de varianza a una vía con efectos fijos se tiene que es el mejor puesto que tiene varianza mínima. ■

Resultado 11.2.4. *Bajo el modelo 11.2.6, el mejor predictor lineal insesgado de T_y y su error cuadrático medio están dados por*

$$\hat{T}_y = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{k \in s_h} Y_{hk} \tag{11.2.9}$$

si se consideran todas las posibles muestras de tamaño n del soporte Q inducido por el diseño de muestreo y se construyen intervalos de 95 % de confianza para la media, entonces se espera que el 95 % de esos intervalos contengan al parámetro μ. Por otra parte, el enfoque inferencial basado en modelos predictivos se debe interpretar en términos del modelo 11.2.1. De esta forma, el procedimiento induce dos variables aleatorias LS y LI tales que $Pr(LI \leq \mu \leq LS) = 0.95$.

$$ECM_\xi(\hat{T}_y) = \sum_{h=1}^{H} \frac{N_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right)\sigma_h^2 \tag{11.2.10}$$

respectivamente.

Prueba. En primer lugar nótese que la variable aleatoria total T se puede reescribir como

$$T_y = \sum_{h=1}^{H}\sum_{k \in U_h} Y_{hk} = \sum_{h=1}^{H} T_{yh}$$

con T_{yh} denotando la variable aleatoria total del estrato h. Recurriendo al resultado 11.2.2 se tiene que $\hat{T}_{yh} = \frac{N_h}{n_h}\sum_{k\in s_h} Y_{hk}$ es un predictor insesgado para T_{yh}. Por tanto

$$E_\xi(\hat{T}_y - T_y) = \sum_{h=1}^{H} E_\xi\left(\hat{T}_{yh} - T_{yh}\right) = 0$$

Por último,

$$\begin{aligned}
ECM_\xi(\hat{T}_y) &= E_\xi\left(\sum_{h=1}^{H}(\hat{T}_{yh} - T_{yh})\right)^2 \\
&= E_\xi\left(\sum_{h=1}^{H}(\hat{T}_{yh} - T_{yh})^2 + \sum_{h}\sum_{g\neq h} E_\xi(\hat{T}_{yh} - T_{yh})(\hat{T}_{yg} - T_{yg})\right) \\
&= E_\xi\left(\sum_{h=1}^{H}(\hat{T}_{yh} - T_{yh})^2\right) = \sum_{h=1}^{H} E_\xi(\hat{T}_{yh} - T_{yh})^2 \\
&= \sum_{h=1}^{H} \frac{N_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right)\sigma_h^2
\end{aligned}$$

■

Análogamente con el modelo para muestreo aleatorio simple, es posible estimar σ_h^2 con S_h^2 en cuyo caso se obtendrían las mimas estimaciones en los dos enfoques.

11.2.3 Un modelo para el muestreo por conglomerados

Suponga que $Y_1, \ldots, Y_N$ es una población de variables aleatorias que se encuentran agrupadas en N_I conglomerados que inducen una partición de la población y al mismo tiempo definen una población de conglomerados $U_I = \{U_1, \ldots, U_{N_I}\}$. El tamaño del i-ésimo conglomerado es N_i $(i = 1, \ldots, N_I)$. El tamaño poblacional general es $N = N_1 + \cdots, N_{N_I}$. El mecanismo probabilístico que rige a la población está dado por un modelo de superpoblación ξ definido como

$$Y_{ik} = \beta + \varepsilon_{ik} \tag{11.2.11}$$

Donde el subscrito ik hace referencia a las cantidades asociadas con el k-ésimo elemento dentro del i-ésimo conglomerado. Cada uno de los ε_{ik} son variables aleatorias independientes e idénticamente distribuidas con media cero y varianza constante σ_i^2 dentro del mismo conglomerado i-ésimo ($i = 1, \ldots, N_I$), con estructura de auto-correlación $\sigma_i\rho_i$ para los elementos pertenecientes al mismo conglomerado i-ésimo y no correlacionados entre conglomerados, tales que

$$\begin{aligned} E_\xi(Y_{ik}) &= \beta \\ Var_\xi(Y_{ik}) &= \sigma_i^2 \\ Cov_\xi(Y_{ik}, Y_{jl}) &= \sigma_i^2\rho_i \quad \text{si } i \neq j \text{ y } k \neq l. \end{aligned} \tag{11.2.12}$$

El modelo indica que todos los elementos tienen una media común. Dentro de los conglomerados, los elementos tienen una varianza común (que puede ser distinta de un conglomerado a otro) y dentro del mismo conglomerado, los elementos comparten un factor de correlación. De esta forma, se selecciona una muestra de conglomerados s_I de tamaño n_I y se observan todos y cada uno de los elementos pertenecientes al conglomerado.

Resultado 11.2.5. *Bajo el modelo 11.2.11, el mejor estimador lineal insesgado de β está dado por*

$$\hat{\beta} = \sum_{i \in s_I} v_i \bar{Y}_{U_i} \tag{11.2.13}$$

donde

$$v_i = \frac{\left(N_i/\sigma_i^2[1 + (N_i - 1)\rho_i]\right)}{\sum_{i \in S_I} (N_i/\sigma_i^2[1 + (N_i - 1)\rho_i])}$$

Prueba. Mediante un argumento similar al de los modelos previos y utilizando el análisis de varianza de efectos aleatorios se tiene la demostración del resultado. ■

Resultado 11.2.6. *Bajo el modelo 11.2.11, el mejor predictor lineal insesgado de T_y y su error cuadrático medio están dados por*

$$\hat{T}_y = \sum_{i \in s_I} \sum_{k=1}^{N_i} Y_{ik} + \sum_{i \notin s_I} N_i \hat{\beta} \tag{11.2.14}$$

$$ECM_\xi(\hat{T}_y) = \sum_{i \notin s_I} N_i \sigma_i^2 [1 + (N_i - 1)\rho_i] + \frac{(N_I - n_I)^2}{\sum_{i \in s_I} N_i/\sigma_i^2[1 + (N_i - 1)\rho_i]} \tag{11.2.15}$$

respectivamente.

Prueba. El lector puede consultar la demostración de este resultado en Royall (1976) y en Scott & Smith (1969) notando que el total puede escribirse como

$$T_y = \sum_{i \in s_I} \sum_{k=1}^{N_i} Y_{ik} + \sum_{i \notin s_I} \sum_{k=1}^{N_i} Y_{ik}$$

■

11.2.4 Un modelo para el muestreo por etapas

Suponga el mismo modelo 11.2.11 pero en esta ocasión se selecciona una muestra de conglomerados s_I de tamaño n_I y para cada conglomerado $U_i \in s_I$ se selecciona una submuestra s_i de tamaño n_i.

Resultado 11.2.7. *Bajo el modelo 11.2.11 y mediante una selección en dos etapas, el mejor estimador lineal insesgado de β está dado por*

$$\hat{\beta} = \sum_{i \in s_I} v_i \bar{Y}_{s_i} \tag{11.2.16}$$

donde

$$v_i = \frac{\left(n_i/\sigma_i^2[1 + (n_i - 1)\rho_i]\right)}{\sum_{i \in S_I} (n_i/\sigma_i^2[1 + (n_i - 1)\rho_i])}$$

Prueba. Mediante un argumento similar al de los modelos previos y utilizando el análisis de varianza anidado con efectos aleatorios se tiene la demostración del resultado. ■

Resultado 11.2.8. *Bajo el modelo 11.2.11 y mediante una selección en dos etapas, el mejor predictor lineal insesgado de T_y está dado por*

$$\hat{T}_y = \sum_{i \in s_I} \sum_{k \in s_i} Y_{ik} + \sum_{i \in s_I} (N_i - n_i) \left[w_i \bar{Y}_{s_i} + (1 - w_i)\hat{\beta}\right] + \sum_{i \notin s_I} N_i \hat{\beta} \tag{11.2.17}$$

con $w_i = n_i\rho_i/[1 + (n_i - 1)\rho_i]$.

Prueba. El lector puede consultar la demostración de este resultado en Royall (1976) y en Scott & Smith (1969) notando que el total puede escribirse como

$$T_y = \sum_{i \in s_I} \sum_{k \in s_i} Y_{ik} + \sum_{i \in s_I} \sum_{k \notin s_i} Y_{ik} + \sum_{i \notin s_I} \sum_{k=1}^{N_i} Y_{ik}$$

■

Nótese que en para el muestreo por conglomerados o por etapas, tanto el predictor como su varianza difieren significativamente del estimador construido mediante el enfoque inferencial basado en el diseño de muestreo.

11.2.5 Un modelo para el estimador de razón

Suponga que $Y_1, \ldots, Y_N$ es una población de variables aleatorias independientes en idénticamente distribuidas y que $X_1, \ldots, X_N$ conforman una población de variables auxiliares tales que su realización para cada uno de los elementos de la población $x_1, \ldots, x_N$ es conocida. El mecanismo probabilístico que rige a la población y que define la relación entre Y_k y X_k está dado por un modelo de superpoblación ξ definido como

$$Y_k = \beta X_k + \varepsilon_k \tag{11.2.18}$$

Donde cada uno de los ε_k $k \in U$ son variables aleatorias independientes e idénticamente distribuidas con media cero y varianza no constante $\sigma^2 X_k$, tales que

$$\begin{aligned} E_\xi(Y_k) &= \beta X_k \\ Var_\xi(Y_k) &= \sigma^2 X_k. \end{aligned} \tag{11.2.19}$$

Este modelo sólo es válido si la línea de regresión pasa por el origen y la varianza se incrementa a medida que la variable auxiliar aumenta su magnitud. De esta población, se selecciona una muestra s de tamaño n. De esta forma se tienen los siguientes resultados.

Resultado 11.2.9. *Bajo el modelo 11.2.18, el mejor estimador lineal insesgado de β está dado por*

$$\hat{\beta} = \frac{\bar{Y}_s}{\bar{X}_s} = \frac{\sum_{k\in s} Y_k}{\sum_{k\in s} X_k} \tag{11.2.20}$$

Prueba. El estimador de β está dado por la minimización de la siguiente función de dispersión

$$D = \sum_{k\in s} \frac{(y_k - \beta)^2}{\sigma^2 X_k}.$$

Luego de derivar e igualar a cero, se encuentra fácilmente que $\hat{\beta} = \dfrac{\sum_{k\in s} Y_k}{\sum_{k\in s} X_k}$. Por otro lado, se tiene que

$$E_\xi(\hat{\beta}) = \frac{1}{\sum_{k\in s} X_k} \sum_{k\in s} E_\xi(Y_k) = \frac{1}{\sum_{k\in s} X_k} \sum_{k\in s} \beta X_k = \beta$$

■

Resultado 11.2.10. *Bajo el modelo 11.2.18, el mejor predictor lineal insesgado de T_y y su error cuadrático medio están dados por*

$$\hat{T}_y = \frac{\bar{Y}_s}{\bar{X}_s} T_x \tag{11.2.21}$$

$$ECM_\xi(\hat{T}_y) = \frac{\sum_{k\notin s} X_k}{\sum_{k\in s} X_k} \sigma^2 T_x \tag{11.2.22}$$

respectivamente, con $T_x = \sum_{k\in U} X_k$.

Prueba. En primer lugar el predictor toma la siguiente forma

$$\hat{T}_y = \sum_{k\in s} Y_k + \sum_{k\notin s} \hat{\beta} X_k = \frac{\bar{Y}_s}{\bar{X}_s} \left[n\bar{X}_s + \sum_{k\notin s} X_k \right] = \frac{\bar{Y}_s}{\bar{X}_s} T_x$$

$\hat{T}_y$ es insesgado puesto que

$$E_\xi(\hat{T}_y - T_y) = E_\xi \left(\frac{\bar{Y}_s}{\bar{X}_s} T_x - \sum_{k\in U} Y_k \right) = \beta T_x - \beta T_x = 0$$

Por último, como el predictor es insesgado, entonces

$$\begin{aligned}
ECM_\xi(\hat{T}_y - T_y) &= Var_\xi(\hat{T}_y - T_y)\\
&= Var_\xi\left(\sum_{k\in s} Y_k + \sum_{k\notin s}\hat{\beta}X_k - \sum_{k\in U} Y_k\right)\\
&= Var_\xi\left(\sum_{k\notin s}\hat{\beta}X_k - \sum_{k\notin s} Y_k\right)\\
&= Var_\xi\left(\sum_{k\notin s}\hat{\beta}X_k\right) + Var_\xi\left(\sum_{k\notin s} Y_k\right)\\
&= \left(\sum_{k\notin s} X_k\right)^2 Var_\xi\left(\frac{\sum_{k\in s} Y_k}{\sum_{k\in s} X_k}\right) + Var_\xi\left(\sum_{k\notin s}(\beta X_k + \varepsilon_k)\right)\\
&= \left(\sum_{k\notin s} X_k\right)^2 \frac{\sum_{k\in s} Var_\xi(Y_k)}{\left(\sum_{k\in s} X_k\right)^2} + Var_\xi\left(\sum_{k\notin s}\varepsilon_k\right)\\
&= \left(\sum_{k\notin s} X_k\right)^2 \frac{\sigma^2\sum_{k\in s} X_k}{\left(\sum_{k\in s} X_k\right)^2} + \sigma^2\sum_{k\notin s} X_k\\
&= \sigma^2\left(\frac{\sum_{k\notin s} X_k}{\sum_{k\in s} X_k}\right) + \left[\sum_{k\notin s} X_k + \sum_{k\in s} X_k\right] = \frac{\sum_{k\notin s} X_k}{\sum_{k\in s} X_k}\sigma^2 T_x
\end{aligned}$$

■

11.2.6 Un modelo para el estimador de regresión

Suponga que $Y_1, \ldots, Y_N$ es una población de variables aleatorias independientes en idénticamente distribuidas y que $X_1, \ldots, X_N$ conforman una población de variables auxiliares tales que su realización para cada uno de los elementos de la población $x_1, \ldots, x_N$ es conocida. El mecanismo probabilístico que rige a la población y que define la relación entre Y_k y X_k está dado por un modelo de superpoblación ξ definido como

$$Y_k = \beta_0 + \beta_1 X_k + \varepsilon_k \tag{11.2.23}$$

Donde cada uno de los ε_k $k \in U$ son variables aleatorias independientes e idénticamente distribuidas con media cero y varianza no constante σ^2, tales que:

$$\begin{aligned}
E_\xi(Y_k) &= \beta_0 + \beta_1 X_k\\
Var_\xi(Y_k) &= \sigma^2.
\end{aligned} \tag{11.2.24}$$

Este modelo sólo es válido si la línea de regresión pasa por el origen y la varianza se incrementa a medida que la variable auxiliar aumenta su magnitud. De esta población, se selecciona una muestra s de tamaño n. De esta forma se tienen los siguientes resultados.

Resultado 11.2.11. *Bajo el modelo 11.2.23, el mejor estimador lineal insesgado de β_0 y β_1 está dado por*

$$\hat{\beta}_1 = \frac{\sum_{k\in s}(x_k - \bar{X}_s)(y_k - \bar{Y}_s)}{\sum_{k\in s}(x_k - \bar{X}_s)^2} \tag{11.2.25}$$

y

$$\hat{\beta}_0 = \bar{Y}_s - \hat{\beta}_1\bar{X}_s \tag{11.2.26}$$

Prueba. Los estimadores se encuentran minimizando la siguiente función de dispersión

$$D = \sum_{k\in s}\frac{(y_k - \beta_0 - \beta_1 X_k)^2}{\sigma^2}.$$

Luego de derivar e igualar a cero, se encuentra fácilmente el resultado. ■

Resultado 11.2.12. *Bajo el modelo 11.2.23, el mejor predictor lineal insesgado de T_y está dado por*

$$\hat{T}_y = N\left(\hat{\beta}_0 + \hat{\beta}_1\bar{X}_U\right) \tag{11.2.27}$$

Prueba. Nótese que el predictor se puede escribir como:

$$\begin{aligned}
\hat{T}_y &= \sum_{k\in s} Y_k + \sum_{k\notin s}(\hat{\beta}_0 + \hat{\beta}_1 X_k)\\
&= n\bar{Y}_s + \sum_{k\notin s}(\hat{\beta}_0 + \hat{\beta}_1 X_k)\\
&= n(\hat{\beta}_0 + \hat{\beta}_1\bar{X}_s) + \sum_{k\notin s}(\hat{\beta}_0 + \hat{\beta}_1 X_k)\\
&= \sum_{k\notin s}(\hat{\beta}_0 + \hat{\beta}_1 X_k) + \sum_{k\notin s}(\hat{\beta}_0 + \hat{\beta}_1 X_k) = \sum_{k\notin U}(\hat{\beta}_0 + \hat{\beta}_1 X_k) = N(\hat{\beta}_0 + \hat{\beta}_1\bar{X}_U)
\end{aligned}$$

■

11.3 El teorema general de predicción

Así como el estimador general de regresión es un caso general de muchos otros estimadores, en el enfoque inferencial basado en los modelos predictivos existe el predictor general de regresión que abarca muchos predictores incluyendo los vistos en la anterior sección. Sin embargo, en esta sección no sólo se estudiarán predicciones de totales poblacionales sino de cualquier función lineal de las variables de interés. El lector notará que el resultado general está basado en la teoría de modelos lineales y en particular el teorema de Gauss-Markov. Aunque en esta sección no hacemos ningún supuesto acerca de distribuciones parametrizadas (como la normal, la gama o la familia exponencial), es posible hacerlo y llegar a resultados óptimos utilizando resultados de inferencia estadística tales como el Lema de Sheffe o el Teorema de Rao-Blackwell (Shao 2003).

Suponga que la población finita consiste de N unidades. El vector de las variables de interés es $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_N)'$ y para cada elemento de la población la

realización de estas variables aleatorias es $\mathbf{y} = (y_1, y_2, \ldots, y_N)'$. Suponga que el objetivo es estimar una combinación lineal[4] $\boldsymbol{\gamma}'\mathbf{y}$. Para tal fin, se selecciona una muestra s de tamaño n. Nótese qué tanto $\mathbf{y}$ cómo $\boldsymbol{\gamma}$ se pueden particionar de la siguiente manera: $\mathbf{y} = (\mathbf{y}_s', \mathbf{y}_r')'$ y $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_s', \boldsymbol{\gamma}_r')'$; en donde el subíndice s se refiere a que el vector contiene los n elementos de la muestra seleccionada y el subíndice r se refiere a que el vector contiene los $N-n$ elementos que no fueron seleccionados en la muestra.

De la anterior manera, es posible reescribir la combinación lineal que se quiere estimar como $\boldsymbol{\gamma}'\mathbf{y} = \boldsymbol{\gamma}_s'\mathbf{y} + \boldsymbol{\gamma}_r'\mathbf{y}_r$, la cual es una realización de la variable aleatoria $\boldsymbol{\gamma}'\mathbf{Y} = \boldsymbol{\gamma}_s'\mathbf{Y} + \boldsymbol{\gamma}_r'\mathbf{Y}_r$. Es claro que el problema de estimar $\boldsymbol{\gamma}'\mathbf{y}$ se reduce al problema de predecir $\boldsymbol{\gamma}_r'\mathbf{y}_r$.

Definición 11.3.1. *Un estimador lineal de* $\theta = \boldsymbol{\gamma}'\mathbf{Y}$ *se define como* $\hat{\theta} = \mathbf{g}_s'\mathbf{Y}_s$. *Donde* $\mathbf{g}_s = (g_1, g_2, \ldots, g_n)'$ *es un vector de tamaño* n.

Definición 11.3.2. *El error de estimación de un estimador* $\hat{\theta}$ *está dado por* $\hat{\theta} - \theta = \mathbf{g}_s'\mathbf{Y}_s - \boldsymbol{\gamma}'\mathbf{Y}$ *y puede ser reescrito como*

$$\begin{aligned}\mathbf{g}_s'\mathbf{Y}_s - \boldsymbol{\gamma}'\mathbf{Y} &= (\mathbf{g}_s'\gamma_s)\mathbf{Y}_s - \boldsymbol{\gamma}_r'\mathbf{Y}_r \\ &= \mathbf{a}'\mathbf{Y}_s - \boldsymbol{\gamma}_r'\mathbf{Y}_r\end{aligned}$$

con $\mathbf{a} = \mathbf{g}_s - \boldsymbol{\gamma}_s$

Nótese que utilizar $\mathbf{g}_s'\mathbf{Y}_s$ para estimar a $\theta = \boldsymbol{\gamma}'\mathbf{Y}$ es equivalente a utilizar $\mathbf{a}'\mathbf{Y}_s$ para predecir $\boldsymbol{\gamma}_r'\mathbf{Y}_r$ y consecuentemente, encontrar un vector óptimo $\mathbf{g}_s$ es equivalente a encontrar un vector óptimo $\mathbf{a}$.

El problema que se aborda en esta sección se enmarca dentro del modelo lineal general dado por

$$\mathbf{Y} = \mathbf{X}_k'\boldsymbol{\beta} + \varepsilon_k \tag{11.3.1}$$

Donde cada uno de los ε_k $k \in U$ son variables aleatorias idénticamente distribuidas con media nula, varianza $Var_\xi(\varepsilon_k) = \sigma_k^2$ y covarianza $Cov_\xi(\varepsilon_k, \varepsilon_l) = \rho_{kl}\sigma_k\sigma_l$ con ρ_{kl} un factor de correlación entre los elementos k y l $(k \neq l)$ tales que:

$$\begin{aligned}E_\xi(Y_k) &= \mathbf{X}_k'\boldsymbol{\beta} \\ Var_\xi(Y_k) &= \sigma_k^2 \\ Cov_\xi(Y_k, Y_l) &= \rho_{kl}\sigma_k\sigma_l \qquad \text{para } k \neq l\end{aligned} \tag{11.3.2}$$

De forma matricial, el anterior modelo queda definido como

$$\begin{aligned}E_\xi(\mathbf{Y}) &= \mathbf{X}\boldsymbol{\beta} \\ Var_\xi(\mathbf{Y}) &= \mathbf{V}\end{aligned} \tag{11.3.3}$$

[4]Si el objetivo es estimar el total poblacional, entonces $\boldsymbol{\gamma}' = (1, 1, \ldots, 1)$. Si el objetivo es estimar la media poblacional, entonces $\boldsymbol{\gamma}' = (1/N, 1/N, \ldots, 1/N)$.

donde X es una matriz de variables auxiliares[5] de tamaño $N \times p$, $\boldsymbol{\beta}$ es un vector de coeficientes de regresión desconocidos de tamaño $p \times 1$ y $\mathbf{V}$ es una matriz de covarianzas definida positiva. Nótese que al momento de seleccionar la muestra, tanto $\mathbf{X}$ como $\mathbf{V}$ pueden ser reescritos como

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_s \\ \mathbf{X}_r \end{pmatrix}, \qquad \mathbf{V} = \begin{pmatrix} \mathbf{V}_{ss} & \mathbf{V}_{sr} \\ \mathbf{V}_{rs} & \mathbf{V}_{rr} \end{pmatrix}$$

donde $\mathbf{X}_s$ es de tamaño $n \times p$, $\mathbf{X}_r$ es de tamaño $(N-n) \times p$, $\mathbf{V}_{ss}$ es de tamaño $n \times n$, $\mathbf{V}_{rr}$ es de tamaño $(N-n) \times (N-n)$, $\mathbf{V}_{sr}$ es de tamaño $n \times (N-n)$ y $\mathbf{V}_{rs} = \mathbf{V}'_{sr}$, asumiendo que $\mathbf{V}_{ss}$ es una matriz definida positiva.

Definición 11.3.3. *Un estimador $\hat{\theta}$ es insesgado si $E_\xi(\hat{\theta}) = 0$*

Definición 11.3.4. *El error cuadrático medio de un estimador $\hat{\theta}$ está dado por $ECM_\xi(\hat{\theta}) = E_\xi(\hat{\theta} - \theta)^2$*

Resultado 11.3.1 (Royall (1976))**.** *El mejor estimador lineal insesgado de θ está dado por*

$$\hat{\theta} = \boldsymbol{\gamma}'_s \mathbf{Y}_s + \boldsymbol{\gamma}'_r \left[\mathbf{X}_r \hat{\boldsymbol{\beta}} + \mathbf{V}_{rs} \mathbf{V}_{ss}^{-1} (\mathbf{Y}_s) \mathbf{X}_s \hat{\boldsymbol{\beta}} \right] \tag{11.3.4}$$

donde

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'_s \mathbf{V}_{ss} \mathbf{X}_s)^{-1} \mathbf{X}'_s \mathbf{V}_{ss} \mathbf{Y}_s \tag{11.3.5}$$

El error cuadrático medio de $\hat{\theta}$ está dado por

$$\begin{aligned} ECM_\xi(\hat{\theta}) &= \boldsymbol{\gamma}'_r (\mathbf{V}_{rr} - \mathbf{V}_{rs} \mathbf{V}_{ss}^{-1} \mathbf{V}_{sr}) \boldsymbol{\gamma}_r \\ &+ \boldsymbol{\gamma}'_r (\mathbf{X}_r - \mathbf{V}_{rs} \mathbf{V}_{ss}^{-1} \mathbf{V}_{sr} \mathbf{X}_s)(\mathbf{X}'_s \mathbf{V}_{ss}^{-1} \mathbf{X}_s)^{-1} (\mathbf{X}_r - \mathbf{V}_{rs} \mathbf{V}_{ss}^{-1} \mathbf{V}_{sr} \mathbf{X}_s)' \boldsymbol{\gamma}_r \end{aligned} \tag{11.3.6}$$

Prueba. En primer lugar, el error cuadrático medio está dado por la siguiente expresión

$$\begin{aligned} E_\xi(\hat{\theta} - \theta)^2 &= E_\xi(\mathbf{a}'\mathbf{Y}_s - \boldsymbol{\gamma}_r \mathbf{Y}_r)^2 \\ &= Var_\xi(\mathbf{a}'\mathbf{Y}_s - \boldsymbol{\gamma}_r \mathbf{Y}_r) + \left(E_\xi(\mathbf{a}'\mathbf{Y}_s - \boldsymbol{\gamma}_r \mathbf{Y}_r)\right)^2 \\ &= \mathbf{a}'\mathbf{V}_{ss}\mathbf{a} - 2\mathbf{a}'\mathbf{V}_{sr}\boldsymbol{\gamma}_r + \boldsymbol{\gamma}'_r \mathbf{V}_{rr} \boldsymbol{\gamma}_r + \left((\mathbf{a}'\mathbf{X}_s - \boldsymbol{\gamma}_r \mathbf{X}_r)\boldsymbol{\beta}\right)^2 \end{aligned}$$

Por un lado, se busca un estimador insesgado, entonces el último sumando debe ser nulo. Es decir $\mathbf{a}\mathbf{X}_s = \boldsymbol{\gamma}'_r \mathbf{X}_r$. Por otro lado, se busca el mejor estimador insesgado; es decir, el estimador de mínimo ECM; esta minimización se hace mediante la técnica de los multiplicadores de Lagrange. Entonces la función a minimizar, restringida al insesgamiento del estimador, es

$$\mathcal{L}(\mathbf{a}, \boldsymbol{\lambda}) = \mathbf{a}'\mathbf{V}_{ss}\mathbf{a} - 2\mathbf{a}'\mathbf{V}_{sr}\boldsymbol{\gamma}_r + \boldsymbol{\gamma}'_r \mathbf{V}_{rr} \boldsymbol{\gamma}_r + 2(\mathbf{a}'\mathbf{X}_s - \boldsymbol{\gamma}_r \mathbf{X}_r)\boldsymbol{\lambda}$$

[5] Se asume que los valores de $\mathbf{X}$ son conocidos para todos los elementos de la población.

donde $\boldsymbol{\lambda}$ es un vector de multiplicadores de Lagrange. Diferenciando con respecto a $\boldsymbol{\lambda}$, $\mathbf{a}$ e igualando a cero, tenemos que

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\lambda}} = \mathbf{a}'\mathbf{X}_s - \boldsymbol{\gamma}_r\mathbf{X}_r = 0$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\lambda}} = 2\mathbf{V}_{ss}\mathbf{a} - 2\mathbf{V}_{sr}\boldsymbol{\gamma}_r + 2\mathbf{X}_s\boldsymbol{\lambda} = 0$$

De la primera ecuación se tiene que

$$\mathbf{a}'\mathbf{X}_s = \boldsymbol{\gamma}_r\mathbf{X}_r \tag{11.3.7}$$

De la segunda ecuación se tiene que

$$\mathbf{a} = \mathbf{V}_{ss}^{-1}(\mathbf{V}_{sr}\boldsymbol{\gamma}_r - \mathbf{X}_s\boldsymbol{\lambda}) \tag{11.3.8}$$

y utilizando la restricción (11.3.7), también se tiene que

$$\boldsymbol{\lambda} = \mathbf{A}_s^{-1}(\mathbf{X}_s'\mathbf{V}_{ss}^{-1}\mathbf{V}_{sr} - \mathbf{X}_r')\boldsymbol{\gamma}_r \tag{11.3.9}$$

con $\mathbf{A}_s = \mathbf{X}_s'\mathbf{V}_{ss}^{-1}\mathbf{X}_s$. Reemplazando esta última expresión en (11.3.8), se encuentra el valor óptimo de $\mathbf{a}$ dado por

$$\mathbf{a}_{opt} = \mathbf{V}_{ss}^{-1}(\mathbf{V}_{sr} - \mathbf{X}_s\mathbf{A}_s^{-1}(\mathbf{X}_s'\mathbf{V}_{ss}^{-1}\mathbf{V}_{sr} - \mathbf{X}_r'))\boldsymbol{\gamma}_r \tag{11.3.10}$$

De esta manera, después de un poco de álgebra, se encuentra que el mejor predictor de $\boldsymbol{\gamma}_r\mathbf{Y}_r$ es

$$\begin{aligned}
\mathbf{a}_{opt}'\mathbf{Y}_s &= \boldsymbol{\gamma}_r'(\mathbf{V}_{rs} - (\mathbf{V}_{rs}\mathbf{V}_{ss}^{-1}\mathbf{X}_s - \mathbf{X}_r)\mathbf{A}_s^{-1}\mathbf{X}_s')\mathbf{V}_{ss}^{-1}\mathbf{Y}_s \\
&= \boldsymbol{\gamma}_r'(\mathbf{V}_{rs}\mathbf{V}_{ss}^{-1}\mathbf{Y}_s - \mathbf{V}_{rs}\mathbf{V}_{ss}^{-1}\mathbf{X}_s\boldsymbol{\beta} + \mathbf{X}_r\boldsymbol{\beta}) \\
&= \boldsymbol{\gamma}_r'(\mathbf{V}_{rs}\mathbf{V}_{ss}^{-1}(\mathbf{Y}_s - \mathbf{X}_s\boldsymbol{\beta}) + \mathbf{X}_r\boldsymbol{\beta})
\end{aligned}$$

De la definición 11.3.1., $\hat{\theta} = \mathbf{g}_s\mathbf{Y}_s$ y de la definición 11.3.2., $\mathbf{g}_s = \mathbf{a} + \boldsymbol{\gamma}_s$. Luego, $\hat{\theta} = \boldsymbol{\gamma}_s'\mathbf{Y}_s + \mathbf{a}'\mathbf{Y}_s$. Reemplazando convenientemente se encuentra la demostración del resultado. El ECM del estimador insesgado está dado por

$$ECM_{\xi}(\hat{\theta}) = \underbrace{\mathbf{a}'\mathbf{V}_{ss}\mathbf{a}}_{P1} - \underbrace{2\mathbf{a}'\mathbf{V}_{sr}\boldsymbol{\gamma}_r}_{P2} + \boldsymbol{\gamma}_r'\mathbf{V}_{rr}\boldsymbol{\gamma}_r$$

Teniendo en cuenta que $\mathbf{A}_s^{-1}\mathbf{X}_s'\mathbf{V}_{ss}^{-1}\mathbf{X}_s = \mathbf{I}$, con $\mathbf{I}$ la matriz identidad y después realizar los pasos algebraicos necesarios se encuentra que la primera parte P1 equivale a

$$\begin{aligned}
P1 &= \boldsymbol{\gamma}_r'\mathbf{V}_{rs}\mathbf{V}_{ss}^{-1}\mathbf{V}_{sr}\boldsymbol{\gamma}_r \\
&\quad - \boldsymbol{\gamma}_r'\mathbf{V}_{rs}\mathbf{V}_{ss}^{-1}\mathbf{X}_s\mathbf{A}_s^{-1}\mathbf{X}_s'\mathbf{V}_{ss}^{-1}\mathbf{V}_{sr}\boldsymbol{\gamma}_r \\
&\quad - \boldsymbol{\gamma}_r'\mathbf{X}_r\mathbf{A}_s^{-1}\mathbf{X}_r'\boldsymbol{\gamma}_r
\end{aligned}$$

y la segunda parte P2 equivale a

$$\begin{aligned}P2 = &\ 2\boldsymbol{\gamma}_r'\mathbf{V}_{rs}\mathbf{V}_{ss}^{-1}\mathbf{V}_{sr}\boldsymbol{\gamma}_r\\ &- 2\boldsymbol{\gamma}_r'\mathbf{V}_{rs}\mathbf{V}_{ss}^{-1}\mathbf{X}_s\mathbf{A}_s^{-1}\mathbf{X}_s'\mathbf{V}_{ss}^{-1}\mathbf{V}_{sr}\boldsymbol{\gamma}_r\\ &+ 2\boldsymbol{\gamma}_r'\mathbf{X}_r\mathbf{A}_s^{-1}\mathbf{X}_s'\mathbf{V}_{ss}^{-1}\mathbf{V}_{sr}\boldsymbol{\gamma}_r\end{aligned}$$

Como el transpuesto de un número es el mismo número, se tiene que sumando las partes necesarias se llega a la demostración completa del teorema. ■

Nótese que todos los estimadores, predictores y modelos de las anteriores secciones son un caso particular de este resultado.

11.4 Ignorando el diseño de muestreo

Gelman, Carlin, Stern & Rubin (2004) explican que se debe ser un estadístico ingenuo si se afirma que toda inferencia debería ser condicional a los datos, sin importar de dónde o cómo fueron seleccionados. Esta es una concepción errada del principio de verosimilitud. La noción de que el método de selección de la muestra es irrelevante en el análisis inferencial puede ser contradicha con un argumento muy simple: suponga que se tienen a disposición diez datos provenientes del lanzamiento de diez dados; todos ellos correspondieron al número seis. La actitud del estadístico acerca de la naturaleza de los datos sería diferente si (1) sólo se hicieron diez lanzamientos, (2) se hicieron sesenta lanzamientos pero se decidió reportar sólo los que resultaron ser seis, (3) apareció diez veces el seis en quinientos lanzamientos y se decidió reportar honestamente estas realizaciones. En tales situaciones es claro que la distribución de los datos observados sigue un patrón completamente distinto que no debe ser ignorado.

En términos generales, un diseño de muestreo no es sino una distribución de probabilidad multivariante definida sobre un conjunto de muestras que pertenecen a un soporte. Pero, una distribución de probabilidad no es más sino un modelo que se asume; en este caso, es un modelo que permite la selección de muestras probabilísticas. Una muestra s induce un vector de inclusión dado por

$$\mathbf{I}(s) = (I_1(s), \ldots, I_k(s), \ldots, I_N(s))'$$

Donde $I_k(s)$ está definida por (2.1.8). Dado el anterior esquema, otra forma de denotar el diseño de muestreo es $f_{\mathbf{I}}(\mathbf{I})$ el cual se conoce para todos los posibles valores de $\mathbf{I}$ en todas las posibles muestras s. Por otro lado, si se asume que la medición de la característica de interés y_k en los individuos de la población está sujeta a un error, entonces éstas deben ser vistas como realizaciones de variables aleatorias Y_k. De esta forma, es necesario definir un modelo para los valores poblacionales que puede depender de cierto parámetro. En este caso, si $Y = (Y_1, \ldots, Y_k, \ldots, Y_N)'$ es el vector poblacional de la característica de interés, entonces $f_{\mathbf{Y}}(\mathbf{Y}; \theta)$ definirá tal modelo.

Para realizar cualquier tipo de inferencias acerca del parámetro θ es necesario trabajar con una distribución de probabilidad conjunta de $(\mathbf{I}, \mathbf{Y})$ que permita unificar todo el esquema anterior en un sólo proceso. La pregunta que atañe al estadístico es la siguiente: ¿cómo se puede expresar esa distribución conjunta en

términos de $f_{\mathbf{I}}(\mathbf{I})$ y de $f_{\mathbf{Y}}(\mathbf{I};\theta)$? Chambers & Skinner (2003) dan la respuesta a esta pregunta motivando la suposición de que $\mathbf{Y}$ sea independiente de $\mathbf{I}$. En algunos caso como en Chambers & Skinner (2003, capítulo 8) el diseño de muestreo depende de los valores de la característica de interés; por ejemplo, en un estudio de casos y controles, la respuesta y_k es de tipo binario, indicando si la k-ésima unidad corresponde a un caso o a un control. A su vez, los casos y controles inducen estratos cuyas muestras son seleccionadas independientemente. En este caso, el diseño de muestreo depende directamente de los valores de la característica de interés. Por lo tanto, la relación entre $\mathbf{I}, \mathbf{Y}$ debe ser expresada como

$$f_{\mathbf{I},\mathbf{Y}}(\mathbf{I},\mathbf{Y};\theta) = f_{\mathbf{I}|\mathbf{Y}}(\mathbf{I}|\mathbf{Y})f_{\mathbf{Y}}(\mathbf{Y};\theta)$$

En este caso, se dice que el diseño de muestreo es *informativo* y no puede ser ignorado en términos de inferencia para θ. Por otro lado, si el diseño de muestreo es *no informativo*, la relación entre $\mathbf{I}, \mathbf{Y}$ debe ser expresada como

$$f_{\mathbf{I},\mathbf{Y}}(\mathbf{I},\mathbf{Y};\theta) = f_{\mathbf{I}}(\mathbf{I})f_{\mathbf{Y}}(\mathbf{Y};\theta)$$

y claramente, el diseño de muestreo puede ser ignorado. Chambers & Skinner (2003) afirman que los diseños de muestreo que dependen directamente de la variable de interés no son raros en la práctica. Sin embargo, los diseños de muestreo implementados cuando el marco de muestreo es muy deficiente como el muestreo en dos fases, en donde se selecciona una primera muestra y con base en los resultados de esta se diseña la estrategia para una segunda submuestra, no puede ser catalogado como no informativo y, por tanto, no puede ser ignorado. Por otro lado, es más común encontrar que el diseño de muestreo dependa de otras variables de información auxiliar, como en el diseño estratificado o el diseño proporcional al tamaño. A continuación se presenta el marco general dado por Valliant, Dorfman & Royall (2000) para modelar conjuntamente el diseño de muestreo y el mecanismo probabilístico que origina a la variable de interés.

Suponga que el diseño de muestreo depende de la variable de interés $\mathbf{Y}$, de algunas variables de información auxiliar reunidas en una matriz $\mathbf{X}$ y de algún vector de parámetros $\boldsymbol{\phi}$, entonces se reescribe como:

$$f_{\mathbf{I}|\mathbf{X},\mathbf{Y}}(\mathbf{I}|\mathbf{X},\mathbf{Y};\boldsymbol{\phi}). \tag{11.4.1}$$

A su vez, la distribución de probabilidad de $\mathbf{Y}$ depende de $\mathbf{X}$ y su relación está supeditada a un vector de parámetros $\boldsymbol{\beta}$, entonces se reescribe como

$$f_{\mathbf{Y}|\mathbf{X}}(\mathbf{Y}|\mathbf{X};\boldsymbol{\beta}). \tag{11.4.2}$$

Resultado 11.4.1. *Un modelo para* $\mathbf{I}, \mathbf{Y}$, *está dado por*

$$f_{\mathbf{I},\mathbf{Y}|\mathbf{X}}(\mathbf{I},\mathbf{Y}|\mathbf{X};\boldsymbol{\phi},\boldsymbol{\beta}) = f_{\mathbf{Y}|\mathbf{X}}(\mathbf{Y}|\mathbf{X};\boldsymbol{\beta})f_{\mathbf{I}|\mathbf{Y},\mathbf{X}}(\mathbf{I}|\mathbf{Y},\mathbf{X};\boldsymbol{\phi}) \tag{11.4.3}$$

Prueba. Aplicando la definición de distribución conjunta y condicional se tiene

el resultado puesto que

$$
\begin{aligned}
f_{\mathbf{I},\mathbf{Y}|\mathbf{X}}(\mathbf{I},\mathbf{Y}|\mathbf{X};\boldsymbol{\phi},\boldsymbol{\beta}) &= \frac{f_{\mathbf{I},\mathbf{Y},\mathbf{X}}(\mathbf{I},\mathbf{Y},\mathbf{X};\boldsymbol{\phi},\boldsymbol{\beta})}{f_{\mathbf{X}}(\mathbf{X};\boldsymbol{\beta})}\\
&= \frac{f_{\mathbf{I}|\mathbf{Y},\mathbf{X}}(\mathbf{I}|\mathbf{Y}|\mathbf{X};\boldsymbol{\phi})f_{\mathbf{Y},\mathbf{X}}(\mathbf{Y},\mathbf{X};\boldsymbol{\beta})}{f_{\mathbf{Y},\mathbf{X}}(\mathbf{Y},\mathbf{X};\boldsymbol{\beta})/f_{\mathbf{Y}|\mathbf{X}}(\mathbf{Y}|\mathbf{X};\boldsymbol{\beta})}\\
&= f_{\mathbf{Y}|\mathbf{X}}(\mathbf{Y}|\mathbf{X};\boldsymbol{\beta})f_{\mathbf{I}|\mathbf{Y},\mathbf{X}}(\mathbf{I}|\mathbf{Y},\mathbf{X};\boldsymbol{\phi})
\end{aligned}
$$

■

Por supuesto, a menos de que se trate de un censo, nunca vamos a observar todos los elementos del vector $\mathbf{Y}$. Es decir, en cierto modo, el modelo (11.4.3) es inútil en términos de inferencia. Cuando se selecciona una muestra s, inmediatamente el vector $\mathbf{Y}$ queda particionado en $\mathbf{Y}_s, \mathbf{Y}_r$. De esta manera la relación entre $\mathbf{I}, \mathbf{Y}_s$ está dada por el siguiente resultado.

Resultado 11.4.2. *La distribución conjunta de* $\mathbf{I}, \mathbf{Y}_s$ *está dada por*

$$
f_{\mathbf{I},\mathbf{Y}_s|\mathbf{X}}(\mathbf{I},\mathbf{Y}_s|\mathbf{X};\boldsymbol{\phi},\boldsymbol{\beta}) = \int f_{\mathbf{Y}|\mathbf{X}}(\mathbf{Y}_s,\mathbf{Y}_r|\mathbf{X};\boldsymbol{\beta})f_{\mathbf{I}|\mathbf{Y},\mathbf{X}}(\mathbf{I}|\mathbf{Y}_s,\mathbf{Y}_r,\mathbf{X};\boldsymbol{\phi})\ d\mathbf{Y}_r \tag{11.4.4}
$$

Prueba. Esta demostración está basada en la definición de función de densidad conjunta y marginal (Mood, Graybill & Boes 1974, p. 141) la cual afirma que si V y W son dos variables aleatorias con densidad conjunta dada por $f_{V,W}(V,W)$ entonces la densidad marginal de V está dada por $\int f_{V,W}(V,W)dW$. En nuestro contexto condicional, nótese que el vector $\mathbf{Y}$ quedó particionado; por lo tanto aplicando el anterior principio y recurriendo al anterior resultado, se tiene que

$$
\begin{aligned}
f_{\mathbf{I},\mathbf{Y}_s|\mathbf{X}}(\mathbf{I},\mathbf{Y}_s|\mathbf{X};\boldsymbol{\phi},\boldsymbol{\beta}) &= \int f_{\mathbf{I},\mathbf{Y}|\mathbf{X}}(\mathbf{I},\mathbf{Y}_s,\mathbf{Y}_r|\mathbf{X};\boldsymbol{\phi},\boldsymbol{\beta})\ d\mathbf{Y}_r\\
&= \int f_{\mathbf{Y}|\mathbf{X}}(\mathbf{Y}_s,\mathbf{Y}_r|\mathbf{X};\boldsymbol{\beta})f_{\mathbf{I}|\mathbf{Y},\mathbf{X}}(\mathbf{I}|\mathbf{Y}_s,\mathbf{Y}_r,\mathbf{X};\boldsymbol{\phi})\ d\mathbf{Y}_r
\end{aligned}
$$

■

Nótese que si el diseño de muestreo es ignorable, entonces el mecanismo probabilístico que gobierna la selección de la muestra no depende de la configuración de los valores poblacionales de la variable de interés; esto significaría que

$$
f_{\mathbf{I}|\mathbf{Y},\mathbf{X}}(\mathbf{I}|\mathbf{Y},\mathbf{X};\boldsymbol{\phi}) = f_{\mathbf{I}|\mathbf{X}}(\mathbf{I}|\mathbf{X};\boldsymbol{\phi}). \tag{11.4.5}
$$

Si esto llegase a suceder, entonces (11.4.4) quedaría convertida en

$$
\begin{aligned}
f_{\mathbf{I},\mathbf{Y}_s|\mathbf{X}}(\mathbf{I},\mathbf{Y}_s|\mathbf{X};\boldsymbol{\phi},\boldsymbol{\beta}) &= f_{\mathbf{I}|\mathbf{X}}(\mathbf{I}|\mathbf{X};\boldsymbol{\phi})\int f_{\mathbf{Y}|\mathbf{X}}(\mathbf{Y}_s,\mathbf{Y}_r|\mathbf{X};\boldsymbol{\beta})\ d\mathbf{Y}_r && (11.4.6)\\
&= f_{\mathbf{I}|\mathbf{X}}(\mathbf{I}|\mathbf{X};\boldsymbol{\phi})f_{\mathbf{Y}_s|\mathbf{X}}(\mathbf{Y}_s|\mathbf{X};\boldsymbol{\beta}) && (11.4.7)
\end{aligned}
$$

En términos de inferencia estadística para el vector de parámetros $\boldsymbol{\beta}$ se tienen los siguientes comentarios:

1. Nótese que (11.4.7) está compuesta por dos términos que se multiplican así

$$f_{\mathbf{I},\mathbf{Y}_s|\mathbf{X}}(\mathbf{I},\mathbf{Y}_s|\mathbf{X};\boldsymbol{\phi},\boldsymbol{\beta}) = \underbrace{f_{\mathbf{I}|\mathbf{X}}(\mathbf{I}|\mathbf{X};\boldsymbol{\phi})}_{h(\mathbf{X})}\underbrace{f_{\mathbf{Y}_s|\mathbf{X}}(\mathbf{Y}_s|\mathbf{X};\boldsymbol{\beta})}_{g(T;\boldsymbol{\beta})}$$

   Lo anterior implica que si existiese una estadística suficiente T para $\boldsymbol{\beta}$, entonces apelando al Criterio de Factorización de Neyman (Mood, Graybill & Boes 1974, p. 306), entonces T estaría contenida en la densidad condicional $f_{\mathbf{Y}_s|\mathbf{X}}(\mathbf{Y}_s|\mathbf{X};\boldsymbol{\beta})$. Razón por la cual, en términos de inferencia estadística para $\boldsymbol{\beta}$, la distribución $f_{\mathbf{I}|\mathbf{X}}(\mathbf{I}|\mathbf{X};\boldsymbol{\phi})$ no contendría ninguna información.

2. Una medida de qué tan bien los datos soportan un parámetro $\boldsymbol{\beta}_2$ comparado con un parámetro $\boldsymbol{\beta}_1$ es el Criterio de Razón de Verosimilitudes (Mood, Graybill & Boes 1974, p. 419) que está dado por

$$\begin{aligned}\frac{f_{\mathbf{I},\mathbf{Y}_s|\mathbf{X}}(\mathbf{I},\mathbf{Y}_s|\mathbf{X};\boldsymbol{\phi},\boldsymbol{\beta}_2)}{f_{\mathbf{I},\mathbf{Y}_s|\mathbf{X}}(\mathbf{I},\mathbf{Y}_s|\mathbf{X};\boldsymbol{\phi},\boldsymbol{\beta}_1)} &= \frac{f_{\mathbf{I}|\mathbf{X}}(\mathbf{I}|\mathbf{X};\boldsymbol{\phi})f_{\mathbf{Y}_s|\mathbf{X}}(\mathbf{Y}_s|\mathbf{X};\boldsymbol{\beta}_2)}{f_{\mathbf{I}|\mathbf{X}}(\mathbf{I}|\mathbf{X};\boldsymbol{\phi})f_{\mathbf{Y}_s|\mathbf{X}}(\mathbf{Y}_s|\mathbf{X};\boldsymbol{\beta}_1)}\\ &= \frac{f_{\mathbf{Y}_s|\mathbf{X}}(\mathbf{Y}_s|\mathbf{X};\boldsymbol{\beta}_2)}{f_{\mathbf{Y}_s|\mathbf{X}}(\mathbf{Y}_s|\mathbf{X};\boldsymbol{\beta}_1)}\end{aligned}$$

   Una vez más, en términos de inferencia para $\boldsymbol{\beta}$, la distribución $f_{\mathbf{I}|\mathbf{X}}(\mathbf{I}|\mathbf{X};\boldsymbol{\phi})$ no contendría ninguna información.

Los anteriores argumentos apuntan a que es posible no tener en cuenta la distribución $f_{\mathbf{I}|\mathbf{X}}(\mathbf{I}|\mathbf{X};\boldsymbol{\phi})$. Si esto llegase a suceder entonces, si un diseño de muestreo es ignorable, se tiene que (11.4.7) quedaría convertida en

$$f_{\mathbf{I},\mathbf{Y}_s|\mathbf{X}}(\mathbf{I},\mathbf{Y}_s|\mathbf{X};\boldsymbol{\phi},\boldsymbol{\beta}) = f_{\mathbf{Y}_s|\mathbf{X}}(\mathbf{Y}_s|\mathbf{X};\boldsymbol{\beta}) \tag{11.4.8}$$

con lo cual se concluye que verdaderamente el mecanismo de selección de la muestra puede ser pasado por alto. Sudgen & Smith (1984) afirman que diseños de muestreo como el aleatorio simple, aleatorio estratificado, proporcional al tamaño, el muestreo a conveniencia o el muestreo balanceado corresponde a casos en donde es posible ignorar el mecanismo de selección. También concluyen que aunque algunas veces los diseños de muestreo pueden ser ignorados en términos de inferencia para $\boldsymbol{\beta}$, es equivocado pensar que siempre pueden ser ignorados en términos de inferencia predictiva para el total poblacional T_y.

Como conclusión, la escogencia del tipo de enfoque (basado en el diseño de muestreo o basado en modelos predictivos) debería estar basada en la adecuación del modelo a la población. Es decir, si el modelo asumido es correcto. Entonces, si se escogió el enfoque basado en modelos predictivos y el modelo no es correcto, entonces las estimaciones estarán sesgadas de la realidad. Por otro lado, las estimaciones que se basan en el diseño de muestreo son robustas e insesgadas a cualquier modelo.

11.5 Ejercicios

11.1 Suponga que el siguiente modelo ξ se ajusta a la población:

$$Y_k = \mu + \sqrt{X_k}\varepsilon_k \qquad k = 1, \ldots, N.$$

donde $E_\xi(\varepsilon_k) = 0$, $Var_\xi(\varepsilon_k) = \sigma^2$ y los errores se consideran independientes. suponga los siguientes predictores para el total poblacional T_Y:

$$\hat{T}_{Y,1} = N\frac{\sum_{k\in s} Y_k/X_k}{\sum_{k\in s} 1/X_k}$$

$$\hat{T}_{Y,2} = \sum_{k\in s} Y_k + (N-n)\frac{\sum_{k\notin s} Y_k/X_k}{\sum_{k\notin s} 1/X_k}$$

(a) Demuestre que tanto $\hat{T}_{Y,1}$ como $\hat{T}_{Y,2}$ son insesgados para el modelo predictivo, de tal forma que $E_\xi(\hat{T}_Y - T_Y) = 0$.

(b) Asumiendo que la muestra s fue seleccionada mediante un diseño de muestreo aleatorio simple, muestre que ninguno de los anteriores predictores es insesgado con respecto a este diseño de muestreo.

(c) Suponga que la muestra s fue seleccionada mediante un diseño de muestreo πPT, con $\pi_k = nX_k/T_X$. Muestre que $\hat{T}_{Y,1}$ es insesgado con respecto a este diseño de muestreo, suponiendo que el tamaño de muestra es grande como para afirmar que la esperanza del cociente es aproximadamente igual al cociente de las esperanzas.

11.2 Genere una población normal de $N = 40$ unidades con media $2X_k$ y varianza 4, con X_k variando entre 10 y 20. Seleccione 10 muestras aleatorias simples de tamaño $n = 5$ de esta población. Calcule $\bar{X}_s$ y $\hat{T}_0 = N\bar{Y}_s$ para estas 10 muestras. ¿Existe alguna correspondencia entre $\bar{X}_s - \bar{X}_U$ y $\hat{T}_0 - T_Y$?

11.3 Suponga que se seleccionó una muestra de tamaño $n = 10$ de una población de $N = 393$ hospitales. La característica de interés es el número de pacientes atendidos en un periodo específico de tiempo. Además se tiene conocimiento de una característica de información auxiliar que corresponde al número de camas de los hospitales. En la población de hospitales, el número total de camas asciende a 107956 y el número total de pacientes atendidos es 320159. Asuma que los valores recolectados son los siguientes:

Y:	41	92	297	377	95	231	601	1063	1645	1894
X:	15	25	80	96	111	125	242	275	551	937

(a) Realice un diagrama de dispersión de Y contra X.

(b) Suponga que se quiere ajustar un modelo de la forma $Y_k = \beta X_k + \varepsilon_k$, con $\varepsilon_k \sim (0, \sigma^2 X_k)$. Calcule el mejor estimador para β y grafique la línea de regresión estimada sobre el diagrama de dispersión de la parte (a).

(c) Suponga que se quiere ajustar un modelo de la forma $Y_k = \beta_0+\beta_1 X_k+\varepsilon_k$, con $\varepsilon_k \sim (0, \sigma^2)$. Calcule el mejor estimador para $\boldsymbol{\beta} = (\beta_0, \beta_1)$ y grafique la línea de regresión estimada sobre el diagrama de dispersión de la parte (a).

(d) Calcule el predictor de expansión $\hat{T}_0 = \frac{N}{n}\sum_s Y_k$, el predictor de razón $\hat{T}_r = N\bar{Y}_s \frac{\bar{X}_U}{\bar{X}_s}$ y el predictor de regresión lineal $\hat{T}_{lr} = \frac{N}{n}\sum_s Y_k + (T_X - \frac{N}{n}\sum_s X_k)\hat{\beta}_1$. Calcule el error de estimación para cada una de las estimaciones. ¿cuál es el efecto de utilizar la característica de información auxiliar?

11.4 Suponga que se seleccionó una muestra de tamaño $n = 10$ de una población de $N = 393$ hospitales. La característica de interés es el número de pacientes atendidos en un periodo específico de tiempo. Además se tiene conociemiento de una característica de información auxiliar que corresponde al número de camas de los hospitales. En la población de hospitales, el número total de camas asciende a 107956 y el número total de pacientes atendidos es 320159. Asuma que los valores recolectados son los siguientes:

Y:	78	315	594	778	410	754	1166	1632	1547	2818
X:	38	70	113	156	227	279	347	437	549	860

(a) Realice un diagrama de dispersión de Y contra X.

(b) Suponga que se quiere ajustar un modelo de la forma $Y_k = \beta X_k + \varepsilon_k$, con $\varepsilon_k \sim (0, \sigma^2 X_k)$. Calcule el mejor estimador para β y grafique la línea de regresión estimada sobre el diagrama de dispersión de la parte (a).

(c) Suponga que se quiere ajustar un modelo de la forma $Y_k = \beta_0+\beta_1 X_k+\varepsilon_k$, con $\varepsilon_k \sim (0, \sigma^2)$. Calcule el mejor estimador para $\boldsymbol{\beta} = (\beta_0, \beta_1)$ y grafique la línea de regresión estimada sobre el diagrama de dispersión de la parte (a).

(d) Calcule el predictor de expansión $\hat{T}_0 = \frac{N}{n}\sum_s Y_k$, el predictor de razón $\hat{T}_r = N\bar{Y}_s \frac{\bar{X}_U}{\bar{X}_s}$ y el predictor de regresión lineal $\hat{T}_{lr} = \frac{N}{n}\sum_s Y_k + (T_X - \frac{N}{n}\sum_s X_k)\hat{\beta}_1$. Calcule el error de estimación para cada una de las estimaciones. ¿cuál es el efecto de utilizar la característica de información auxiliar?

11.5 Escriba el siguiente programa de simulación:

(a) Genere una población de tamaño $N = 200$ que sigue un modelo $Y_k = X_k + \varepsilon_k\sqrt{X_k}$, con ε_k independientes con distribución normal estándar y X_k variando entre 10 y 20.

(b) Seleccione 50 muestras aleatorias simples sin reemplazo de tamaño $n = 30$, y calcule el predictor de razón para cada muestra.

(c) Calcule el predictor de regresión bajo el modelo $Y_k = X_k + \varepsilon_k$.

(d) Compare, empíricamente, los sesgos y los errores cuadrados medios de los predictor de regresión y de razón.

(e) ¿Cuánta eficiencia se pierde al utilizar especificaciones incorrectas de la varianza?

11.6 Considere las siguientes situaciones hipotéticas para algunos estudios:

(a) Se selecciona una muestra aleatoria estratificada de hospitales mediante el uso de la estratificación de la población por tipos de servicio que provee cada hospital. Se necesita realizar estimaciones acerca del tiempo promedio de permanencia por paciente, clasificado por tipo de enfermedad para un trimestre particular. Dado que los hospitales más grandes de la región no tienden a responder, entonces aún si son seleccionados aleatoriamente, éstos no se tendrán en cuenta.

(b) Suponga que en la parte (a), se intenta entrevistar a todos los hospitales seleccionados (tanto los grandes como los demás), pero la mitad de los hospitales seleccionados se rehúsan a responder la entrevista.

(c) Suponga que en la parte (a), todos los hospitales seleccionados aceptan responder la entrevista, pero los hospitales grandes sólo proveen información durante la tercera semana de cada mes.

Discuta si los anteriores mecanismos de muestreo son ignorables o no.

11.7 Demuestre que el predictor general de regresión $\hat{T}_{Y,greg} = \hat{T}_{Y,\pi} + (\mathbf{T_X} - \hat{\mathbf{T}}_{\mathbf{X},\pi})'\hat{\boldsymbol{\beta}}$ es insesgado bajo el modelo $Y_k = \mathbf{X}_k'\boldsymbol{\beta} + \varepsilon_k$. Suponga que los errores tienen media cero y varianza constante.

Parte III

Otros tópicos avanzados de muestreo

Capítulo 12

Muestreo en dos fases

> Existen numerosos ejemplos que muestran cómo la información auxiliar disponible puede ser usada [en la estrategia de muestreo] para lograr mayor precisión en las estimaciones. Sin embargo, si la información auxiliar no está disponible pero se conoce que puede ser recolectada de forma barata y en gran escala, valdría la pena conseguir tal información en una primera fase y luego seleccionar una muestra para la característica de interés.
>
> Des Raj (1968)

12.1 Introducción

Propuesto por Neyman (1938), el muestreo en dos fases es el diseño indicado cuando no se tienen conocimiento pleno del comportamiento estructural de la población de interés y esto se ve reflejado en un pésimo marco de muestreo que no contempla variables de información auxiliar (de ningún tipo: ni discreto, ni continuo) y por ello, no es posible proponer el uso de una estrategia de muestreo óptima (diseños avanzados proporcionales al tamaño o estratificados y estimadores de regresión o de calibración) para la estimación de los parámetros poblacionales de interés.

En (Särndal & Sweensson 1987) aparece un marco general que desarrolla la teoría del muestreo en dos fases de modo teórico e inducido por los principios del estimador de Horvitz-Thompson. El diseño de muestreo en dos fases (también conocido como muestreo bifásico o muestreo doble) se utiliza cuando existe poco o nulo conocimiento sobre el comportamiento de la característica de interés a través de los individuos que conforman la población. Por ejemplo, el estimador de razón combinada requiere que todos los elementos de la población puedan ser estratificados y que el total poblacional de la característica de información auxiliar, $t_x = \sum_U x_k$, sea conocido; sin embargo, en muchos casos prácticos no se tiene este tipo de información auxiliar (pertenencia de los miembros de la población a estratos específicos o el total poblacional de las características de información auxiliar). En estos casos en donde el marco de muestreo contiene poca o deficiente información para proponer un diseño de muestreo eficiente, el estadístico puede recurrir a las siguientes

dos opciones (Särndal, Swensson & Wretman 1992):

1. Usar un diseño de muestreo simple como el muestreo aleatorios simple sin reemplazo o el muestreo aleatorio de conglomerados y combinarlo con el estimador de Horvitz-Thompson para ganar más precisión conforme el tamaño de muestra aumenta.

2. Obtener información acerca de la población para construir un nuevo marco muestral. Si se utiliza el estimador de regresión se logra una precisión deseada con un tamaño de muestra moderado.

Nótese que la asignación de un tamaño de muestra grande o la construcción de un nuevo marco muestral implican el desgaste de recursos económicos y logísticos que tal vez el estudio no pueda soportar. De esta manera, una tercera opción es usar un diseño de muestreo en dos fases:

a) En la primera fase, se selecciona una muestra de tamaño n_a - moderado, más no pequeño - de elementos, la cual será denotada como S_a. La selección de esta primera muestra se realiza mediante un diseño $p_a(\cdot)$. Para cada uno de los elementos en S_a se debe obtener información sobre una o más variables auxiliares[1]. Esta muestra queda determinada por las variables aleatorias

$$I_k = \begin{cases} 1, & \text{si el elemento } k \text{ está en la muestra de la primera fase} \\ 0, & \text{si el elemento } k \text{ no está en la muestra de la primera fase} \end{cases}$$

Por lo tanto la probabilidad de inclusión de un elemento en la primera muestra S_a de la primera fase está dada por la siguiente expresión

$$\pi_{ak} = Pr(I_k = 1) = \sum_{s_a \ni k} p_a(s_a) \qquad (12.1.1)$$

y la probabilidad de inclusión de segundo orden en S_a está dada por

$$\pi_{akl} = Pr(I_k I_l = 1) = \sum_{S_a \ni \ k \text{ y } l} p_a(s_a) \qquad (12.1.2)$$

b) En la segunda fase, con la ayuda de la información obtenida en la primera fase, se selecciona una submuestra S de tamaño n, de S_a, mediante un diseño de muestreo $p(\cdot \mid s_a)$. A continuación se observa la característica de interés para los elementos seleccionados en la submuestra. Esta muestra queda determinada por las variables aleatorias

$$D_k = \begin{cases} 1, & \text{si el elemento } k \text{ está en la muestra de la segunda fase} \\ 0 & \text{si el elemento } k \text{ no está en la muestra de la segunda fase} \end{cases}$$

[1]Nótese que este proceso resulta menos costoso que obtener la información directamente de la población.

La probabilidad de que un elemento esté en esta submuestra depende de lo que haya pasado en la primera fase. La probabilidad de inclusión de los elementos en la muestra de la segunda fase está dada por la siguiente expresión

$$\pi_{k|s_a} = Pr(D_k = 1 \mid \mathbf{I}) = \sum_{s \ni k} p(s|s_a) \tag{12.1.3}$$

donde $\mathbf{I} = (I_1, \ldots, I_N)'$ denota el vector de inclusión de la primera muestra. Por otro lado, la probabilidad de inclusión de segundo orden en S está dada por

$$\pi_{kl|s_a} = Pr(D_k D_l = 1 \mid \mathbf{I}) = \sum_{S \ni \ k \text{ y } l} p(s|s_a) \tag{12.1.4}$$

Por ejemplo, Lohr (2000) afirma que en una encuesta de empresas se podría extraer una muestra, en la primera fase, de declaraciones de impuestos y registrar el ingreso reportado por cada empresa seleccionada en esta primera fase (esta muestra puede ser grande puesto que se asume que no es costoso obtener la información auxiliar). En una segunda fase, se podría pensar en seleccionar una submuestra con probabilidad proporcional al ingreso medido en la primera fase, o bien, utilizar la información del ingreso para estratificar las empresas de la muestra de la primera fase y luego establecer contacto con un subconjunto de empresas en cada estrato con el fin de obtener la información deseada acerca de características de interés como gastos totales o impuestos declarados.

El autor recalca que el diseño de muestreo que proporciona el soporte de muestreo que contempla tanto la primera como la segunda fase, no está dado por $p_a(s_a)$ ni por $p(s|s_a)$ sino que, recurriendo al teorema de probabilidad total (Mood, Graybill & Boes 1974), está dado por la siguiente expresión

$$p(s) = \sum_{s_a \supset s} p_a(s_a) p(s|s_a) \tag{12.1.5}$$

Y por lo tanto la probabilidad de inclusión de cualquier elemento en la muestra final S, es

$$\begin{aligned}
\pi_k = Pr(I_k D_k = 1) &= \sum_{s \ni k} \sum_{s_a \supset s} p_a(s_a) p(s|s_a) \\
&= \sum_{s_a \ni k} \sum_{\substack{S_a \subset S \\ s \ni k}} p_a(s_a) p(s|s_a) \\
&= \sum_{s_a \ni k} p_a(s_a) \sum_{\substack{S_a \subset S \\ s \ni k}} p(s|s_a) \\
&= \sum_{s_a \ni k} p_a(s_a) \pi_{k|s_a}
\end{aligned} \tag{12.1.6}$$

Por lo tanto, bajo este tipo de esquemas de muestreo en dos fases, no es posible utilizar los principios del estimador de Horvitz-Thompson, en términos de inferencia del total poblacional, puesto que aunque es posible conocer el valor de las probabilidades inducidas por $p_a(s_a)$ para cada muestra S_a, no es posible conocer siempre los valores de las probabilidades de inclusión en la segunda fase $\pi_{k|s_a}$ para cada muestra S_a puesto que éstos están supeditados a la realización de la primera muestra.

12.2 El estimador π^*

Nótese que otro posible estimador del total poblacional de la características de interés es $\sum_{S_a} y_k/\pi_{ak}$, este es otro estimador inútil puesto que sólo se podría calcular si y_k y π_{ak} fueran conocidos para todo $k \in s_a$. Pero y_k solamente es conocido en la submuestra para $k \in s$. Por lo tanto, condicional a s_a, la muestra de la primera fase, la siguiente cantidad, $\sum_{s_a} y_k/\pi_{ak}$, es estimada insesgadamente por el estimador de Horvitz-Thompson condicionado mediante

$$\hat{t}_{y,\pi^*} = \sum_s \frac{y_k}{\pi^*_k} = \sum_s \frac{y_k}{\pi_{ak}\pi_{k|s_a}} \tag{12.2.1}$$

y definido como el estimador π^* (Särndal & Sweensson 1987).

Resultado 12.2.1. *En muestreo bifásico el total poblacional t_y es estimado insesgadamente por el estimador π^*. Además la varianza del estimador y la estimación insesgada de la varianza están dadas por*

$$Var_{Bif}(\hat{t}_{y,\pi^*}) = \sum\sum_U \Delta_{akl}\frac{y_k}{\pi_{ak}}\frac{y_l}{\pi_{al}} + E_{p_a}\left(\sum\sum_{S_a}\Delta_{kl|S_a}\frac{y_k}{\pi^*_k}\frac{y_l}{\pi^*_l}\right) \tag{12.2.2}$$

$$\widehat{Var}_{Bif}(\hat{t}_{y,\pi^*}) = \sum\sum_S \frac{\Delta_{akl}}{\pi^*_{kl}}\frac{y_k}{\pi_{ak}}\frac{y_l}{\pi_{al}} + \sum\sum_S \frac{\Delta_{kl|S_a}}{\pi_{kl|S_a}}\frac{y_k}{\pi^*_k}\frac{y_l}{\pi^*_l} \tag{12.2.3}$$

respectivamente, con $\pi^*_k = \pi_{ak}\pi_{k|S_a}$, $\pi^*_{kl} = \pi_{akl}\pi_{kl|S_a}$, $\Delta_{akl} = \pi_{akl} - \pi_{ak}\pi_{al}$ y $\Delta_{kl|S_a} = \pi_{kl|S_a} - \pi_{k|S_a}\pi_{l|S_a}$, donde cada sumando de (12.2.3) es insesgado para su contraparte en (12.2.2).

Prueba. Al usar el condicionamiento sucesivo del resultado 7.1.3, para la estructura probabilística del diseño de muestreo p_a, se tiene que

$$\begin{aligned}
E_{Bif}(\hat{t}_{y,\pi^*}) &= E_{p_a}\left(E_p\left(\hat{t}_{y,\pi^*} \mid \mathbf{I}\right)\right)\\
&= E_{p_a}\left(E_p\left(\sum_s \frac{y_k}{\pi^*_k} \mid \mathbf{I}\right)\right)\\
&= E_{p_a}\left(\sum_{S_a} E_p(D_k \mid \mathbf{I})\frac{y_k}{\pi_{ak}\pi_{k|s_a}}\right)\\
&= \sum_U E_{p_a}(I_k)\frac{y_k}{\pi_{ak}} = \sum_U y_k = t_y
\end{aligned}$$

Para probar los resultados de la varianza se utiliza un razonamiento similar dado que

$$Var_{Bif}(\hat{t}_{y,\pi^*}) = Var_{p_a}(E_p(\hat{t}_{y,\pi^*}|\mathbf{I})) + E_{p_a}(Var_p(\hat{t}_{y,\pi^*}|\mathbf{I}))$$

Para el primer sumando se tiene que, utilizando los principios del estimador de

Horvitz-Thompson

$$
\begin{aligned}
Var_{p_a}(E_p(\hat{t}_{y,\pi^*}|\mathbf{I})) &= Var_{p_a}\left(E_p\left(\sum_s \frac{y_k}{\pi_k^*} \mid \mathbf{I}\right)\right)\\
&= Var_{p_a}\left(\sum_{S_a} \frac{y_k}{\pi_{ak}}\right)\\
&= \sum\sum_U \Delta_{akl}\frac{y_k}{\pi_{ak}}\frac{y_l}{\pi_{al}}
\end{aligned}
$$

Para el segundo sumando se procede similarmente, haciendo $y_{ak} = y_k/\pi_{ak}$ se tiene que

$$
\begin{aligned}
E_{p_a}(Var_p(\hat{t}_{y,\pi^*}|\mathbf{I})) &= E_{p_a}\left(Var_p\left(\sum_s \frac{y_k}{\pi_k^*} \mid \mathbf{I}\right)\right)\\
&= E_{p_a}\left(Var_p\left(\sum_s \frac{y_{ak}}{\pi_{k|S_a}} \mid \mathbf{I}\right)\right)\\
&= E_{p_a}\left(\sum\sum_{S_a} \Delta_{kl|S_a}\frac{y_{ak}}{\pi_{k|S_a}}\frac{y_{al}}{\pi_{l|S_a}}\right)\\
&= E_{p_a}\left(\sum\sum_{S_a} \Delta_{kl|S_a}\frac{y_k}{\pi_k^*}\frac{y_l}{\pi_l^*}\right)
\end{aligned}
$$

Por otro lado notando que $E(D_kD_l|\mathbf{I}) = \pi_{kl|S_a}$ y $E(I_kI_l) = \pi_{akl}$ se tiene el insesgamiento de la estimación de la varianza. ■

Ejemplo 12.2.1. Continuando con nuestra población ejemplo U de tamaño $N = 5$, suponga que en una primera fase se selecciona una muestra de $n_a = 2$ elementos de acuerdo a un diseño de muestreo aleatorio simple. En la segunda fase se selecciona una submuestra de $n = 1$ de acuerdo a un diseño de muestreo aleatorio simple[2].

Para la primera fase, y recurriendo al ejemplo 2.1.1, las $\binom{N}{n_a}$ posibles muestras, junto con su respectiva probabilidad de selección, son

	X1	X2	p_a
1	Yves	Ken	0.1
2	Yves	Erik	0.1
3	Yves	Sharon	0.1
4	Yves	Leslie	0.1
5	Ken	Erik	0.1
6	Ken	Sharon	0.1
7	Ken	Leslie	0.1
8	Erik	Sharon	0.1
9	Erik	Leslie	0.1
10	Sharon	Leslie	0.1

[2] Aunque utilizar en las dos fases un diseño de muestreo aleatorio simple no es realista en la vida práctica, este ejemplo sirve para tener una mayor comprensión acerca de la estructura probabilística inducida por el muestreo en dos fases.

La probabilidad de inclusión en la muestra de la primera fase, para cada uno de los 5 elementos de U, es

$$\pi_{ak} = \frac{n_a}{N} = \frac{2}{5}$$

Para la segunda fase existen $\binom{n}{n_a}$ posibles submuestras por cada muestra de la primera fase, el diseño de muestreo de la segunda fase y el diseño de muestreo general queda definido de la siguiente manera

```
        X1       X2     p_a     S       p( |s_a)     p(s)

1      Yves      Ken    0.1    Yves          0.5     0.05
                               Ken           0.5     0.05

2      Yves     Erik    0.1    Yves          0.5     0.05
                               Erik          0.5     0.05

3      Yves   Sharon    0.1    Yves          0.5     0.05
                               Sharon        0.5     0.05

4      Yves   Leslie    0.1    Yves          0.5     0.05
                               Leslie        0.5     0.05

5       Ken     Erik    0.1    Ken           0.5     0.05
                               Erik          0.5     0.05

6       Ken   Sharon    0.1    Ken           0.5     0.05
                               Sharon        0.5     0.05

7       Ken   Leslie    0.1    Ken           0.5     0.05
                               Leslie        0.5     0.05

8      Erik   Sharon    0.1    Erik          0.5     0.05
                               Sharon        0.5     0.05

9      Erik   Leslie    0.1    Erik          0.5     0.05
                               Leslie        0.5     0.05

10   Sharon   Leslie    0.1    Sharon        0.5     0.05
                               Leslie        0.5     0.05
```

Nótese que, recurriendo al teorema de probabilidad total, el diseño de muestreo final, que contempla la dinámica probabilística de la primera y segunda fase, queda definido como sigue a continuación:

$$p(s) = \begin{cases} 0.2, & \text{si } s = \{\textbf{Yves}\}, \\ 0.2, & \text{si } s = \{\textbf{Ken}\}, \\ 0.2, & \text{si } s = \{\textbf{Erik}\}, \\ 0.2, & \text{si } s = \{\textbf{Sharon}\}, \\ 0.2, & \text{si } s = \{\textbf{Leslie}\}. \end{cases}$$

La probabilidad de inclusión de un elemento de S_a en la submuestra de la última fase, condicionada a la realización de una muestra particular, está dada por

$$\pi_{k|S_a} = \frac{n_a}{n} = \frac{1}{2}$$

Luego la probabilidad de inclusión de un elemento de U condicional dada por π_k^* es

$$\pi_k^* = \pi_{ak}\pi_{k|S_a} = \frac{n_a}{N}\frac{n_a}{n} = \frac{n}{N} = \frac{1}{5}$$

que, para este caso particular coincide con la probabilidad de inclusión (propiamente dicha) del elemento dada en (12.1.6). Sin embargo, casi siempre $\pi_k^* \neq \pi_k$ como se demuestra con la siguiente configuración inducida por un diseño de muestreo con probabilidades de selección desiguales.

	X1	X2	p_a	S	p(|S_a)	p(s)
1	Yves	Ken	0.25	Yves	0.9	0.225
				Ken	0.1	0.025
2	Yves	Erik	0.15	Yves	0.8	0.120
				Erik	0.2	0.030
3	Yves	Sharon	0.15	Yves	0.7	0.105
				Sharon	0.3	0.045
4	Yves	Leslie	0.10	Yves	0.6	0.060
				Leslie	0.4	0.040
5	Ken	Erik	0.10	Ken	0.5	0.050
				Erik	0.5	0.050
6	Ken	Sharon	0.05	Ken	0.4	0.020
				Sharon	0.6	0.030
7	Ken	Leslie	0.05	Ken	0.3	0.015
				Leslie	0.7	0.035
8	Erik	Sharon	0.05	Erik	0.2	0.010
				Sharon	0.8	0.040
9	Erik	Leslie	0.05	Erik	0.1	0.005
				Leslie	0.9	0.045
10	Sharon	Leslie	0.05	Sharon	0.5	0.025
				Leslie	0.5	0.025

Nótese que, para esta configuración, y una vez más recurriendo al teorema de probabilidad total, el diseño de muestreo final, queda definido de la siguiente manera:

$$p(s)=\begin{cases}0.510, & \text{si } s=\{\textbf{Yves}\},\\ 0.110, & \text{si } s=\{\textbf{Ken}\},\\ 0.140, & \text{si } s=\{\textbf{Sharon}\},\\ 0.095, & \text{si } s=\{\textbf{Erik}\},\\ 0.145, & \text{si } s=\{\textbf{Leslie}\}.\end{cases}$$

En este caso, para la primera fase, la probabilidad de inclusión en la muestra de la primera fase, para cada uno de los 5 elementos de U, es

$$\pi_{ak}=\begin{cases}0.65, & \text{si } k=\textbf{Yves},\\ 0.45, & \text{si } k=\textbf{Ken},\\ 0.35, & \text{si } k=\textbf{Erik},\\ 0.30, & \text{si } k=\textbf{Sharon},\\ 0.25, & \text{si } k=\textbf{Leslie}.\end{cases}$$

La probabilidad de inclusión de un elemento de S_a en la submuestra de la segunda fase, condicionada a la realización de una muestra particular, está dada por los siguientes 10 casos (tantos casos como muestras en la primera fase)

- Si $S_a=S_1$, entonces

$$\pi_{k|S_a}=\begin{cases}0.90, & \text{si } k=\textbf{Yves},\\ 0.10, & \text{si } k=\textbf{Ken}.\end{cases}$$

- Si $S_a=S_2$, entonces

$$\pi_{k|S_a}=\begin{cases}0.80, & \text{si } k=\textbf{Yves},\\ 0.20, & \text{si } k=\textbf{Erik}.\end{cases}$$

- Y así sucesivamente, hasta
- Si $S_a=S_{10}$, entonces

$$\pi_{k|S_a}=\begin{cases}0.50, & \text{si } k=\textbf{Sharon},\\ 0.50, & \text{si } k=\textbf{Leslie}.\end{cases}$$

Por lo tanto, también existirán 10 casos para el cálculo de la cantidad π_k^*, así:

- Si $S_a=S_1$, entonces

$$\pi_k^*=\begin{cases}0.65\times 0.90=0.585, & \text{si } k=\textbf{Yves},\\ 0.45\times 0.10=0.045, & \text{si } k=\textbf{Ken}.\end{cases}$$

- Si $S_a = S_2$, entonces

$$\pi_k^* = \begin{cases} 0.65 \times 0.80 = 0.520, & \text{si } k = \textbf{Yves}, \\ 0.35 \times 0.20 = 0.007, & \text{si } k = \textbf{Erik}. \end{cases}$$

- Y así sucesivamente, hasta
- Si $S_a = S_{10}$, entonces

$$\pi_k^* = \begin{cases} 0.30 \times 0.50 = 0.150, & \text{si } k = \textbf{Sharon}, \\ 0.25 \times 0.50 = 0.125, & \text{si } k = \textbf{Leslie}. \end{cases}$$

Lo anterior muestra que $\pi_k^* \neq \pi_k$, puesto que la probabilidad de inclusión está dada por

$$\pi_k = \begin{cases} 0.510, & \text{si } k = \textbf{Yves}, \\ 0.110, & \text{si } k = \textbf{Ken}, \\ 0.140, & \text{si } k = \textbf{Erik}, \\ 0.095, & \text{si } k = \textbf{Sharon}, \\ 0.145, & \text{si } k = \textbf{Leslie}. \end{cases}$$

Nótese que en la vida práctica, con poblaciones bastante grandes, no es posible calcular π_k. Como ejercicio, utilizando los datos del ejemplo 2.1.3, se debe corroborar el insesgamiento del estimador π_k^* tanto en la primera como en esta última configuración.

12.3 Estratificación en muestreo bifásico

Hidiroglou & Rao (2003) afirman que la primera propuesta de Neyman (1938) fue la estratificación en muestreo bifásico, en donde en la primera fase se selecciona una muestra aleatoria S_a de tamaño n_a. El siguiente paso es observar una variable de información auxiliar x_k para cada elemento $k \in S_a$ y con base en el comportamiento de esta característica se estratifica la muestra S_a; es decir todo elemento $k \in S_a$ se clasifica en un y sólo un estrato h con $h = 1 \ldots, H$, de tal forma que

$$S_a = \bigcup_{h=1}^{H} S_{ah} \qquad n_a = \sum_{h=1}^{H} n_{ah}$$

en donde S_{ah} corresponde al h-ésimo estrato de tamaño n_{ah}, que comúnmente se considera aleatorio. En la segunda fase se selecciona una muestra S_h de tamaño fijo[3] n_h para cada estrato $h = 1, \ldots, H$, de tal forma que

$$S = \bigcup_{h=1}^{H} S_h \qquad n = \sum_{h=1}^{H} n_h$$

[3]Hidiroglou & Rao (2003) afirman que el supuesto de que n_h es fijo es inconsistente puesto que depende de la variable n_{ah}, la cual varia de cero hasta $\text{mín}(n_1, N_h)$, donde N_h corresponde al tamaño poblacional del estrato h.

en donde S corresponde a la submuestra de la segunda fase de tamaño n. Nótese que la muestra de la primera fase S_a se selecciona mediante un diseño arbitrario $p_a(s_a)$ mientras que la submuestra de la segunda fase S_h dentro de cada estrato $h = 1, \ldots, H$ también se selecciona mediante un diseño arbitrario en cada estrato[4] denotado por $p_h(S_h|S_a)$.

Resultado 12.3.1. *Bajo este marco de referencia, el total poblacional t_y es estimado insesgadamente por*

$$\hat{t}_{y,\pi^*} \sum_{h=1}^{H} \sum_{S_h} \frac{Y_k}{\pi_k^*} \tag{12.3.1}$$

Además, la varianza del estimador y la estimación insesgada de la varianza están dadas por

$$\begin{aligned} Var_{Bif}(\hat{t}_{y,\pi^*}) &= \sum\sum_{U} \Delta_{akl} \frac{y_k}{\pi_{ak}} \frac{y_l}{\pi_{al}} \\ &\quad + E_{p_a}\left(\sum_{h=1}^{H} \sum\sum_{S_{ah}} \Delta_{kl|S_a} \frac{y_k}{\pi_k^*} \frac{y_l}{\pi_l^*} \right) \end{aligned} \tag{12.3.2}$$

$$\widehat{Var}_{Bif}(\hat{t}_{y,\pi^*}) = \sum\sum_{S} \frac{\Delta_{akl}}{\pi_{kl}^*} \frac{y_k}{\pi_{ak}} \frac{y_l}{\pi_{al}} + \sum_{h=1}^{H} \sum\sum_{S_h} \frac{\Delta_{kl|S_a}}{\pi_{kl|S_a}} \frac{y_k}{\pi_k^*} \frac{y_l}{\pi_l^*} \tag{12.3.3}$$

respectivamente, donde cada sumando de (12.3.3) es insesgado para su contraparte en (12.3.2).

Suponga que, en la primera fase, se extrae una muestra aleatoria simple S_a de tamaño n_a de una población de tamaño N. Por tanto,

$$\pi_{ak} = \frac{n_a}{N} \qquad \pi_{akl} = \frac{n_a(n_a-1)}{N(N-1)} \tag{12.3.4}$$

Luego, con la información recopilada en la primera fase, es posible separar las unidades en H estratos distintos (sólo se sabe a qué estrato pertenece el elemento hasta que se selecciona la muestra en la primera fase). Luego, para cada estrato, mediante un diseño de muestreo aleatorio simple, se selecciona una muestra de tamaño n_h, suponiendo que los estratos son de tamaño n_{ah} con $h = 1, 2, ..., H$. Luego, para la segunda fase, la probabilidad de inclusión de un elemento está dado por

$$\pi_{k|s_a} = \frac{n_h}{n_{ah}} \qquad \text{para } k \in S_{ah} \text{ con } h = 1, \ldots, H \tag{12.3.5}$$

y la probabilidad de inclusión de segundo orden es

$$\pi_{kl|s_a} = \begin{cases} \frac{n_h}{n_{ah}} & \text{si } k = l \in S_{ah} \\ \frac{n_h(n_h-1)}{n_{ah}(n_{ah}-1)} & \text{si } k \neq l,\ k, l \in S_{ah} \\ \frac{n_h}{n_{ah}} \frac{n_{h'}}{n_{ah'}} & \text{si } k \in S_{ah},\ l \in S_{ah'} \end{cases} \tag{12.3.6}$$

[4]La propuesta inicial de Neyman (1938) fue utilizar un diseño aleatorio simple tanto para la selección de la primera muestra en la primera fase como para la selección de las submuestras de la segunda fase en cada estrato.

De lo anterior se tiene que el estimador del total poblacional es

$$\hat{t}_{y,\pi^*} = \sum_S \frac{y_k}{\pi_k^*} = \frac{N}{n_a} \sum_{S_h} \frac{n_{ah}}{n_h} y_k \tag{12.3.7}$$

Para calcular la varianza se procede con el condicionamiento sucesivo de la siguiente manera

$$\begin{aligned}
Var_{Bif}(\hat{t}_{y,\pi^*}) &= Var_{MAS}(E_{MAE}(\hat{t}_{y,\pi^*} \mid \mathbf{I})) + E_{MAS}(Var_{MAE}(\hat{t}_{y,\pi^*} \mid \mathbf{I})) \\
&= Var_{MAS}\left(\frac{N}{n_a}\sum_{S_a} y_k\right) \\
&= +E_{MAS}\left(Var_{MAE}\left(\frac{N}{n_a}\sum_{S_h}\frac{n_{ah}}{n_h} y_k \mid \mathbf{I}\right)\right) \\
&= \underbrace{\frac{N^2}{n_a}\left(1-\frac{n_a}{N}\right) S^2_{y_U}}_{V_1} + \underbrace{\frac{N^2}{n_a^2} E_{MAS}\left(\sum_{h=1}^{H} \frac{n_{ah}^2}{n_h}\left(1-\frac{n_h}{n_{ah}}\right) S^2_{y_{ah}}\right)}_{V_2}
\end{aligned}$$

donde el primer término hace referencia a la varianza de la muestra en la primera fase mientras que el segundo término hace referencia a la varianza adicional debida al submuestreo en la segunda fase. Nótese que $S^2_{y_{ah}}$ es la varianza de la característica de interés en el estrato h-ésimo de la muestra de la primera fase. Es importante recalcar que en el segundo término, el operador E_{MAS} está especificado sobre todas y cada una de las posibles muestras estratificadas de la segunda fase.

Rao (1973) propuso la estimación para estos componentes de varianza los cuales son estimados insesgadamente por las siguientes expresiones

$$\hat{V}_1 = \frac{N^2}{n_a}\left(1-\frac{n_a}{N}\right)\sum_{h=1}^{H}\frac{n_{ah}}{n_a}\left\{(1-Q_h)S^2_{y_{S_h}} + \frac{n_a}{n_a-1}(\bar{y}_{S_h}-\bar{y}_S)\right\}$$

$$\hat{V}_2 = \frac{N^2}{n_a^2}\left(\sum_{h=1}^{H}\frac{n_{ah}^2}{n_h}\left(1-\frac{n_h}{n_{ah}}\right)S^2_{y_{ah}}\right)$$

respectivamente, y donde $Q_h = \dfrac{(n_a - n_{ah}}{n_h(n_a-1)}$. La demostración de este resultado puede ser consultada en Hidiroglou & Rao (2003).

12.4 Selección proporcional al tamaño

En las secciones anteriores se ha podido comprobar cómo la información auxiliar puede ser usada para ganar precisión y eficiencia en la estimación del total de una característica de interés. En algunas ocasiones esta información puede ser utilizada en la etapa de diseño y en otras en la etapa de estimación. Cuando se

quiere utilizarla en la etapa de diseño se puede utilizar un diseño de muestreo proporcional a alguna característica de información auxiliar x. En esta ocasión se presentará la segunda opción.

Si se sabe que el comportamiento estructura de la característica de información auxiliar es proporcional al comportamiento de la característica de interés, entonces sería deseable seleccionar la muestra con probabilidad proporcional a x. Sin embargo, esta información x no está disponible a nivel poblacional, pero se sabe que es barato conseguirla al menos en una muestra grande. Por tanto, ésta se recolecta en una muestra inicial s_a de tamaño n_a inducida por un diseño de muestreo aleatorio simple de una población de tamaño N. Después de que sea posible tener acceso a esta información auxiliar, entonces se selecciona una submuestra s de tamaño m con reemplazo proporcional a la variable de información auxiliar x.

Resultado 12.4.1. *Bajo este marco de referencia en donde la muestra inicial s_a de tamaño n_a es seleccionada mediante muestreo aleatorio simple y la submuestra s de tamaño m es seleccionada proporcional a x, entonces el estimador insesgado del total poblacional, su varianza y su varianza estimada están dados por*

$$\hat{t}_y = \frac{N}{n_a}\hat{t}_{ay} = \frac{N}{n_a}\frac{1}{m}\sum_{k\in S}\frac{y_k}{p_{ak}} = \frac{N}{n_a}\frac{t_{ax}}{m}\sum_{k\in S}\frac{y_k}{x_k} \tag{12.4.1}$$

$$\begin{aligned} Var_{Bif}(\hat{t}_y) &= \frac{N^2}{n_a}\left(1-\frac{n_a}{N}\right)S^2_{y_U} \\ &\quad + \frac{N(n_a-1)}{(N-1)n_a}\frac{1}{m}\sum_U \frac{1}{p_k}\left(\frac{y_k}{p_{ak}}-t_y\right)^2 \end{aligned} \tag{12.4.2}$$

$$\begin{aligned} \widehat{Var}_{Bif}(\hat{t}_y) &= \frac{N^2}{n}\frac{t^2_{ax}}{m(m-1)}\left[\sum_{k\in S}\frac{y_k^2}{x_k^2}-\frac{1}{m}\left(\sum_{k\in S}\frac{y_k}{x_k}\right)^2\right] \\ &\quad + \frac{N(N-n_a)}{mn_a(n_a-1)}\left(t_{ax}\sum_{k\in S}\frac{y_k^2}{x_k}+\frac{t^2_{ax}}{n_a(m-1)}\left[\sum_{k\in S}\frac{y_k^2}{x_k^2}-\frac{1}{m}\left(\sum_{k\in S}\frac{y_k}{x_k}\right)^2\right]\right) \end{aligned} \tag{12.4.3}$$

respectivamente, con $\hat{t}_{ay} = \frac{1}{m}\sum_s \frac{y_k}{p_{ak}}$, $p_{ak} = \frac{x_k}{t_{ax}}$ y $t_{ax} = \sum_{S_a} x_k$.

Prueba. Utilizando una vez más la propiedad del condicionamiento sucesivo se tiene que

$$\begin{aligned} E(\hat{t}_y) &= E_{MAS}\left(\frac{N}{n}E_{PPT}\left(\sum_s \frac{y_k}{p_{ak}}|\mathbf{I}\right)\right) \\ &= E_MAS\left(\frac{N}{n}\sum_{s_a} y_k\right) = t_y \end{aligned}$$

Y concerniente al primer término de la varianza se tiene que

$$Var_{MAS}(E_{PPT}(\hat{t}_y)) = Var_{MAS}\left(\frac{N}{n_a}\sum_{s_a} y_k\right) = \frac{N^2}{n_a}\left(1-\frac{n_a}{N}\right)S^2_{y_U}$$

Para el segundo término, acudiendo al resultado 2.2.14 y al resultado 4.2.6, nótese que

$$
\begin{aligned}
Var_{PPT}(\hat{t}_y|\mathbf{I}) &= \frac{N^2}{n_a^2} Var_{PPT}\left(\frac{1}{m}\sum_{s}\frac{y_k}{p_k}|\mathbf{I}\right)\\
&= \frac{N^2}{n_a^2}\frac{1}{m}\sum_{k\in S_a} p_{ak}\left(\frac{y_k}{p_{ak}} - t_{ay}\right)^2 = \frac{N^2}{n_a^2}\frac{1}{m}\sum_{S_a}\sum_{k<l} p_k p_l \left(\frac{y_k}{p_k} - \frac{y_l}{p_l}\right)^2
\end{aligned}
$$

Por lo tanto, se tiene que

$$
\begin{aligned}
E_{MAS}(Var_{PPT}(\hat{t}_y)) &= E_{MAS}\left(\frac{N^2}{n_a^2}\frac{1}{m}\sum_{S_a}\sum_{k<l} p_k p_l \left(\frac{y_k}{p_k} - \frac{y_l}{p_l}\right)^2\right)\\
&= E_{MAS}\left(\frac{N^2}{n_a^2}\frac{1}{m}\sum_{U}\sum_{k<l} p_k p_l \left(\frac{y_k}{p_k} - \frac{y_l}{p_l}\right)^2 I_k I_l\right)\\
&= \frac{N^2}{n_a^2}\frac{1}{m}\sum_{U}\sum_{k<l} p_k p_l \left(\frac{y_k}{p_k} - \frac{y_l}{p_l}\right)^2 E_{MAS}(I_k I_l)\\
&= \frac{N^2 n_a(n_a-1)}{n_a^2 N(N-1)}\frac{1}{m}\sum_{U}\sum_{k<l} p_k p_l \left(\frac{y_k}{p_k} - \frac{y_l}{p_l}\right)^2\\
&= \frac{N(n_a-1)}{(N-1)n_a}\frac{1}{m}\sum_{U}\frac{1}{p_k}\left(\frac{y_k}{p_{ak}} - t_y\right)^2
\end{aligned}
$$

Lo anterior usando la forma alternativa de la varianza del diseño de muestreo PPT. La demostración de la estimación insesgada de la varianza del estimador puede ser consultada en Raj (1968, p.143). ■

12.5 Otras aplicaciones

Este diseño de muestreo bifásico tiene muchas aplicaciones en la vida práctica y los tópicos que se han tocado hasta el momento no son sino una breve introducción al complejo y basto mundo de las encuestas por muestreo con sus deficiencias y limitaciones. Sin embargo, este capítulo ha mostrado que sí es posible afrontar estas limitaciones desde el punto de vista teórico y encontrar una solución mediática a estos problemas. A continuación, un breve resumen de otras aplicaciones del muestreo bifásico.

12.5.1 Mejorando el estimador

Este capítulo se enfocó en la búsqueda de un diseño de muestreo óptimo y en el mejoramiento de la forma de selección de muestras en la segunda etapa. Sin embargo, es posible considerar un diseño de muestreo muy simple y sencilla en ambas etapas pero con la ayuda de información auxiliar, recopilada en la muestra de la primera fase, mejorar el estimador al utilizar el planteamiento del estimador general de regresión o de los estimadores de calibración. Por supuesto, dependiendo

de la calidad de la información conseguida, es posible mejorar tanto el diseño de muestreo como el estimador.

Como lo afirma Estevao & Särndal (2001), una característica distintiva del muestreo en dos fases es que la información auxiliar puede ser encontrada en varios niveles:

- A nivel poblacional completo: el valor de cada una de las características de información auxiliar se conoce para todos y cada uno de los individuos que pertenecen a la población.
- A nivel poblacional incompleto: sólo se conoce el valor de los totales de las características de información auxiliar mas no se valor individual.
- A nivel de la primera fase S_a: el valor de cada una de las características de información auxiliar se conoce para todos y cada uno de los individuos que pertenecen a la muestra de la primera fase S_a.
- A nivel de la segunda fase S: el valor de cada una de las características de información auxiliar se conoce para todos y cada uno de los individuos que pertenecen a la submuestra de la segunda fase S.

Alguna información reposa en el nivel poblacional mientras que otra lo hace en el nivel de la muestra en la primera fase de muestreo. Aun teniendo acceso a las dos, el investigador decide a discreción si utiliza ambas o alguna o incluso ninguna para obtener estimaciones eficientes. La varianza del estimador (de regresión o de calibración) dependerá entonces del nivel en que se encuentre la información auxiliar que se ha decido utilizar. Es importante identificar cuál es el tipo de información auxiliar que es relevante para el estudio puesto que no siempre es posible encontrar la información auxiliar completa; pero inclusio si es posible encontrarla, se debe definir si se va a utilizar o no; puesto que

1. En algunas situaciones, la eficiencia puede decrecer dramáticamente si se ignora alguna característica de información auxiliar en el proceso de calibración. Incluso es posible obtener un estimador de calibración cuya varianza sea menor que la de aquel construido con base en información auxiliar completa.
2. No siempre es posible contar con información auxiliar completa así que se debe lograr el objetivo de mejorar la estimación con la información que se tiene a la mano. Es importante conocer cómo este tipo de limitaciones afecta la varianza del estimador.

Estevao & Särndal (2001) han mostrado que existen exactamente diez casos diferentes conteniendo distintas configuraciones de información auxiliar para los estimadores de calibración y da cuenta de la varianza de los mismos dependiendo del caso. El tratamiento de Särndal & Sweensson (1987) para el estimador general de regresión es exhaustivo y comprende una muy buena fuente de referencia para estrategias de muestreo de tipo bifásico para las cuales en la etapa de estimación consideran un modelo de superpoblación para asistir en la eficiencia del estimador. Esta lectura puede ser complementada con el capítulo 9 de Särndal, Swensson & Wretman (1992).

12.5.2 Un modelo para la ausencia de respuesta

Las personas que no responden con frecuencia difieren de manera crucial de las personas que sí lo hacen. De esta forma, es posible hacer la siguiente clasificación: a) **la ausencia de respuesta por unidad**, en donde falta toda la unidad de observación y suele suceder porque el encuestador no pudo establecer contacto con el hogar, la persona seleccionada está enferma o se rehúsa a participar. En esta etapa el encuestador debe determinar algunas características demográficas del hogar para su posterior imputación y b) **la ausencia de respuesta por registro**, en donde faltan algunos registros de la unidad de observación aunque otros si están efectivamente respondidos. Los siguientes son algunos puntos de vista para enfrentar la ausencia de Respuesta:

- Prevención: diseñar la encuesta de modo que la ausencia de respuesta se pequeña. Éste es el mejor método de enfrentarla.
- Sub-muestra: seleccionar una sub-muestra representativa de las unidades que no respondieron y realizar inferencias.
- Modelos: utilizar un modelo para predecir los valores de las unidades que no respondieron. Es decir reemplazar los registros de la unidad faltante, por registros predichos resultantes del modelo.
- Ignorancia: es una práctica muy común ignorar la ausencia de respuesta en la encuesta y realizar inferencias con los datos recopilados de las unidades respondientes.

La ausencia de repuesta conlleva grandes efectos[5] en los resultados de calidad de las estimaciones. Por ejemplo, si se aumentara el tamaño de muestra para enfrentar la ausencia de respuesta, es posible que nos encontremos con una mayor cantidad de personas de la misma clase de respondientes (homogeneidad). Nótese que el sesgo puede aumentar porque se malgastaron recursos que hubiesen servido para remediar la ausencia de respuesta. Por otro lado, si se omite el efecto de la ausencia de respuesta en una encuesta de victimización, se subestima el número total de víctimas. Ahora, en la población se forman dos estratos «respondientes» y «no respondientes» y el sesgo se reduce si el promedio es similar en los dos estratos (esta opción es imposible de conocer pues los «no respondientes» simplemente no responden) o si hay poca ausencia de respuesta.

Lohr (2000) plantea que algunos de los factores que inciden en el aumento de la ausencia de respuesta pueden ser:

1. Contenido: encuestas relacionadas con el uso de drogas, finanzas. Se puede acotar la tasa de respuesta si se ordenan las preguntas de manera adecuada.
2. Tiempo de la encuesta: algunas temporadas arrojan tasas de no respuestas más altas que otras.
3. Encuestadores: aplicar métodos estándar de mejoramiento de la calidad para aumentar la precisión y tasa de respuesta de los entrevistadores involucrados en el estudio.

[5] Si se insiste en calcular y estimar totales y medias, sin tener en cuenta la ausencia de respuesta, se debe informar en el reporte técnico la cifra correspondiente a la tasa de respuesta.

4. Método de recolección: las encuestas telefónicas y por correo tienen una tasa de respuesta menor que las entrevistas personales[6].
5. Diseño de cuestionario: formulación de las preguntas.
6. Agobio: encuestas demasiado largas que indisponen al respondiente.
7. Presentación de la encuesta: es el primer contacto entre el respondiente y el encuestador.
8. Incentivos: los incentivos financieros o «regalos» aumentan la tasa de respuesta. Los anti-incentivos también son de utilidad, por ejemplo la suspensión de la licencia de conducción al negarse a contestar.

Brewer (2002) afirma que la ausencia de respuesta y el muestreo en dos fases están relacionados de la siguiente manera: la forma más sencilla de tratar con la ausencia de respuesta es tratando a la muestra de respondientes como si éstos constituyesen la muestra objetivo, o equivalentemente como si la población de respondientes efectivo y no respondientes estuvieran gobernados por la misma estructura de probabilidad. De esta manera, la muestra objetivo es tratada como la muestra de la primera fase y el conjunto de respondientes efectivos es tratada como la submuestra de la segunda fase.

Särndal & Lundström (2004) menciona que este enfoque comienza con el supuesto de que la distribución de las respuestas es conocida (aunque en la práctica no es así). Esto implica que las probabilidades de respuesta de primer y segundo orden están dadas por

$$Pr(k \in r|S) = \theta_k \qquad Pr(k, l \in r|S) = \theta_{kl} \tag{12.5.1}$$

las cuales se asumen conocidas y donde r denota el grupo de respondientes efectivos y S la muestra total conformada por respondientes y no respondientes. De esta forma es posible calcular las ponderaciones combinadas (nótese la similitud con la construcción de la cantidad π_k^*) $(1/\pi_k) \times (1/\theta_k)$ y calcular el siguiente estimador insesgado de dos fases

$$\hat{t}_y = \sum_{k \in r} \frac{y_k}{\pi_k \theta_k} \tag{12.5.2}$$

Como las probabilidades de respuesta θ_k son desconocidas, entonces el anterior estimador es imposible de calcular. Por tanto, para hacerlo operacional, se debe encontrar una estimación de estas. Suponga que existen características de información auxiliar disponibles que permiten obtener un estimador (o también predictor) de esta probabilidad, denotado como $\hat{\theta}_k$. Por lo tanto, se ha obtenido un estimador de dos fases que contempla la ausencia de respuesta reemplazando θ_k por $\hat{\theta}_k$ y dado por

$$\hat{t}_y = \sum_{k \in r} \frac{y_k}{\pi_k \hat{\theta}_k} \tag{12.5.3}$$

Existen distintas formas de encontrar estimadores $\hat{\theta}_k$, algunos de ellos son discutidos en el capítulo 9 de Särndal, Swensson & Wretman (1992).

[6]Utilizar un sistema CATI (entrevista telefónica asistida por computador, por sus siglas en inglés) mejora la precisión de los datos.

12.5.3 Muestreo en ocasiones

En muchos estudios de investigación se seleccionan muestras de la misma población de manera repetida en el tiempo y la misma característica de interés se mide en cada ocasión. De esta manera, el comportamiento estructural de ésta puede ser medido a través del tiempo. El muestreo en dos ocasiones considera una población finita y en la primera ocasión, se selecciona una muestra S_a mediante un diseño de muestreo $p_a(\cdot)$ y se mide la característica de interés y. En la segunda ocasión se seleccionan dos muestras independientes, una muestra traslapada, S_t, proveniente de la anterior muestra S_a y otra no traslapada, S_{nt} tomada del complemento de la primera muestra S_a^c. En el capítulo 9 de Särndal, Swensson & Wretman (1992) se aborda la teoría para el tratamiento de la anterior configuración de muestreo.

12.6 Marco y Lucy

A continuación se utiliza la población de empresas del sector industrial para ejemplificar el desarrollo del muestreo en dos fases y cómo éste permite mejorar bastante la estrategia de muestreo. En esta sección se contemplan tres configuraciones que muestran claramente escenarios difíciles pero comunes en la vida práctica, en donde las encuestas y los marcos de muestreo sufren de imperfecciones y es necesario afilar las herramientas estadísticas para poder tratar con estos problemas.

Primera configuración: estratificación

En este primer escenario se considera que el marco de muestreo es deficiente y sólo contempla la ubicación e identificación de las empresas del sector industrial. Bajo este marco de referencia se supone que no se conoce absolutamente nada acerca del comportamiento estructural de la población a través de las variables de interés: Ingreso, Gastos e Impuestos declarados durante el año pasado.

Suponga que el investigador conoce que el sector industrial está dividido en tres niveles. Grande, Mediano y Pequeño y que además el comportamiento de las características de interés es sustancialmente diferente en cada uno de los anteriores subgrupos poblacionales. Si las bondades del marco de muestreo llegaran hasta determinar la clasificación de cada empresa a alguno de los anteriores tres estratos, entonces podría utilizarse un diseño de muestreo estratificado para mejorar la estimación. Sin embargo, suponga que no es posible contar con tal información a nivel poblacional. Sin embargo, existen algunas entidades de origen privado que venden esta información a un precio razonable. La mala noticia es que, debido a conflictos de intereses, no entregan la lista completa sino un subconjunto de 1000 de las 2396 empresas del sector industrial. La buena noticia es que el investigador puede determinar las mil empresas a su gusto.

Bajo la anterior configuración, es posible utilizar un diseño de muestreo bifásico de la siguiente manera: en la primera fase, seleccionar una muestra de tamaño $n_a = 1000$ y obtener la información del nivel para cada una de las empresas incluidas en esta primera muestra. Para esto, se utiliza la función `S.SI` del paquete `TeachingSampling` para obtener la primera muestra que será llamada como `Fase1`.

```
> data(Lucy)
> N <- dim(Lucy)[1]
> n <- 1000
> sam <- S.SI(N,n)
> Fase1 <- Lucy[sam,]
> attach(Fase1)

> Fase1
          ID   Ubication      Level
2      AB002        c1k2      Small
3      AB003        c1k3      Small
4      AB004        c1k4      Small
...
1577   AB049      c16k92     Medium
1579   AB058      c16k94     Medium
1581   AB076      c16k96     Medium
...
2315   AB098      c25k38        Big
2318  AB1028      c25k41        Big
2320  AB1038      c25k43        Big
...
```

La muestra realizada en la primera fase es de tamaño 1000 y está dividida en cada uno de los tres estratos. Por otro lado, en la segunda fase, y acudiendo a la información de pertenencia a los estratos, se selecciona una segunda muestra estratificada de tamaño $n = 400$ y para esto se configura la función `S.STSI` del paquete `TeachingSampling`.

```
> na1 <- summary(Level)[[1]]
> na2 <- summary(Level)[[2]]
> na3 <- summary(Level)[[3]]
> na1;       na2;         na3
[1] 33    [1] 298     [1] 669
n.a <- c(na1,na2,na3)

> n1 <- 30
> n2 <- 120
> n3 <- 250
> n <- c(n1,n2,n3)

> sam <- S.STSI(Level,n.a,n)
> data <- Fase1[sam,]

> data
         ID Ubication  Level Zone Income Employees Taxes SPAM
510  AB1685      c6k15  Small    B    235        15   2.0   no
1396  AB417    c15k10  Small    C    149        27   0.5  yes
377  AB1564      c4k80  Small    B    300        66   3.0   no
859  AB2010      c9k67  Small    C    451        79   9.0  yes
...
```

```
1998  AB621    c22k18 Medium    D    590         55  16.0  yes
2187  AB846     c24k9 Medium    E    750         51  28.0   no
2002  AB626    c22k22 Medium    D    580        109  16.0   no
...

2390  AB957    c26k14    Big    E   1220        163  63.0   no
2333 AB1123    c25k56    Big    A   1020        133  50.0  yes
2342  AB118    c25k65    Big    A   1118        109  59.0   no
> attach(data)
```

La submuestra realizada en la segunda fase es de tamaño 400 y está dividida en cada uno de los tres estratos. Una vez conseguida la información, se procede a estimar las cantidades de interés. Para esto se utiliza la función `E.STSI` del paquete `TeachingSampling`, la cual arroja las estimaciones expandidas a la muestra de la primera fase. Para expandirlas a la población basta con multiplicarlas por el inverso de la probabilidad de inclusión de la primera muestra[7].

Los resultados se muestran a continuación.

```
> estima <- data.frame(Income, Employees, Taxes)
> E.STSI(Level,n.a,n,estima)
> dim(E.STSI(Level,n.a,n,estima))
[1] 3 4 4

> (2396/1000)*E.STSI(Level,n.a,n,estima)[1,,]

                  N     Income Employees     Taxes
Big          79.068   99130.19  11116.96  6024.982
Medium      714.008  472709.00  56912.39 15809.327
Small      1602.924  449870.24  81729.89  6020.583
Population 2396.000 1021709.42 149759.24 27854.891
```

Nótese que esta estrategia es recomendable cuando se desean obtener estimaciones eficiente por subgrupos poblacionales.

Segunda configuración: selección proporcional al tamaño

En este apartado suponga que se tienen las mismas condiciones que en el escenario anterior. Sin embargo, el interés ahora no se centra en la estimación eficiente de los totales de la característica de interés dentro de algunos subgrupos poblacionales sino en la estimación eficiente del total poblacional de las características de interés. De esta manera, se desea ejecutar un diseño de muestreo aleatorio simple, en una primera etapa, para poder incorporar información auxiliar en la segunda etapa. Como antes, se utiliza la función `S.SI` del paquete `TeachingSampling` para la selección de esta primera muestra.

[7]Esta operación **solamente** tiene sentido para las estimaciones de los totales y no para las varianzas ni sus estimaciones. Por lo tanto, estas se deben obviar puesto que no conducen al verdadero valor de las cantidades mencionadas.

```
> library(TeachingSampling)
> data(Lucy)
> N <- dim(Lucy)[1]
> na <- 1000
> sam <- S.SI(N,na)
> Fase1 <- Lucy[sam,]
> attach(Fase1)
```

Una vez se ha seleccionado la muestra, el investigador se ve forzado a recopilar información auxiliar que le permita mejorar la estrategia de muestreo. En este caso, el investigador conoce que la característica Ingreso está relacionada directamente con las características de interés Número de Empleados e Impuestos. Además, es fácil conseguir tal información, puesto que, al igual que en la configuración anterior, existe una entidad que suministra dicha información aunque sólo para 1000 empresas por términos de cláusulas de confidencialidad. De esta manera, el investigador recopila los datos de Ingreso para las 1000 empresas incluidas en la muestra de la primera fase y toma la decisión de mejorar la estrategia de muestreo por medio de la incorporación de esta información auxiliar en el diseño de muestreo. En este orden de ideas, él decide utilizar un diseño de muestreo proporcional al Ingreso de las empresas. Para la selección de la submuestra se utiliza la función `S.PPS` del paquete `TeachingSampling`. La submuestra es de tamaño $m = 400$ y se selecciona con reemplazo.

```
> n <- 400
> res <- S.PPS(400,Income)
> sam <- res[,1]
> pk.s <- res[,2]
> sum(pk.s)
[1] 0.571502
> data <- Fase1[sam,]
> attach(data)
> estima <- data.frame(Income, Employees, Taxes)
```

Para la estimación del total poblacional de las características de interés se procede con la función `E.PPS` del paquete `TeachingSampling`, la cual provee la estimación expandida en la muestra de la Fase 1. Para expandir los resultados a la población, una vez más, basta con multiplicar estos resultados por el inverso de la probabilidad de inclusión de la primera fase dada por 2396/1000.

```
> (2396/1000)*E.PPS(estima,pk.s)[1,]
    Income  Employees      Taxes
1043671.24  148385.29   30030.74
```

Tercera configuración: estimación de calibración

Para este último escenario, suponga que el investigador selecciona una muestra aleatoria simple para la primera fase de muestreo con el fin de recolectar información que le permita mejorar la estrategia de muestreo.

```
> library(TeachingSampling)
> library(sampling)
> data(Lucy)
> N <- dim(Lucy)[1]
> na <- 1000
> sam <- S.SI(N,na)
> Fase1 <- Lucy[sam,]
> attach(Fase1)
```

Suponga ahora, que la entidad que provee la información, está dispuesta a brindar para cada una de las empresas incluidas en la muestra de la primera fase, no sólo la información del Ingreso sino que también la información acerca del Número de Empleados. De esta forma, el investigador propone seleccionar una submuestra mediante un diseño de muestreo aleatorio simple y combinarlo con un estimador de calibración mediante el método de Raking.

```
> t.ax  <- c(1000, sum(Income), sum(Employees))
> n <- 400
> sam <- S.SI(na,n)
> data <- Fase1[sam,]
> attach(data)
```

Para estimar los resultados expandidos a la primera fase se utiliza la función `calib` del paquete `Sampling`, la cual proporciona las ponderaciones calibradas para la Fase 1. De la misma manera, estos resultados se expanden a la población mediante la multiplicación del inverso de la probabilidad de inclusión de la primera muestra.

```
> y.as  <- data.frame(Income, Employees, Taxes)
> x.as  <- cbind(1,Income, Employees)
> pi.ak <- rep(n/na,times=n)
> w.ak <- calib(x.as, d=1/pi.ak, t.ax, method="raking")

> tc.a <- t(w.ak/pi.ak)%*%as.matrix(y.as)
> (2396/1000)*tc.a
      Income Employees    Taxes
[1,] 1015890  150610.2 27994.88
```

Comparación de resultados

Aunque a primera vista, parecería que los resultados no tan cercanos a los totales poblacionales verdaderos, nótese que en particular para la características de interés Ingreso se obtiene una ganancia amplia comparado con un diseño de muestreo aleatorio simple. Nótese también que en este caso, el estimador de calibración arroja mejores resultados.

Tabla 12.1: *Estimaciones realizadas bajo distintos escenarios para el muestreo bifásico.*

Método	Total poblacional	Total estimado	Desv. %
Estratos	28654	27854	-2.79
Proporcional	28654	30031	4.81
Calibración	28654	27995	-2.29

12.7 Ejercicios

12.1 Suponga un estudio longitudinal que plantea tres encuestas, tipo semipanel, en diferentes tiempos. Para la tercera medición, se utilizó un diseño de muestreo con una rotación del 20 % para las siguientes posibles especificaciones:

- De tamaño n_1 que fue seleccionada sólo de la muestra de la primera medición.
- De tamaño n_{12} que fue seleccionada de las muestras de las mediciones uno y dos.
- De tamaño n_{123} que fue seleccionada de las muestras de las tres mediciones.
- De tamaño n_{23} que fue seleccionada de las muestras de las mediciones dos y tres.
- De tamaño n_3 que fue seleccionada de la muestra de la tercera medición.

a. Dibuje un diagrama que ilustre la rotación de la muestra en las tres mediciones y los tamaños relativos de las cinco configuraciones anteriores.

b. Proponga una fórmula para la estimación del total poblacional de la característica de interés en la tercera medición para las cinco configuraciones anteriores.

c. Sin escribir ninguna fórmula estadística para las varianzas, indique en cuál de estas configuraciones y por qué, induce mayor eficiencia en las estimaciones.

12.2 Suponga un diseño de muestreo en dos fases. En la primera fase, se seleccionó una muestra aleatoria simple sin reeemplazo s_a de tamaño $n_a = 150$. En esta fase se levantó la información de una característica de interés x. En la segunda fase, se decidió seleccionar una muestra s, mediante un diseño de muestreo Poisson con tamaño de muestra esperado $n_s = 10$, mediante probabilidades de inclusión proporcionales a la característica de información auxiliar. La información para la muestra de la segunda fase se encuentra consolidada en la tabla 12.2.

a. Calcule una estimación insesgada para el total poblacional de y, teniendo en cuenta que el total de la característica de interés en la muestra de la primera fase es 4060.

b. Utilice la siguiente expresión para calcular el respectivo coeficiente de variación estimado

$$\widehat{Var}(\hat{t}_{y,\pi^*}) = \left(\frac{N}{n}\bar{x}_{s_a}\right)^2 \frac{1}{n_a - 1}\left[(n_a - f_a)\sum_s \left(\frac{y_k}{x_k}\right)^2 - (1 - f_a)\left(\sum_s \frac{y_k}{x_k}\right)^2\right]$$
$$- \frac{N}{n}\bar{x}_{s_a}\sum_s \frac{y_k^2}{x_k}$$

Tabla 12.2: Tabla del ejercicio 12.2

y	x
2653	33
17949	247
1060	12
1324	12
2223	18
2553	30
2216	20
13205	138
3475	35
7072	62
4623	47

12.3 Asuma que la muestra de la segunda fase del ejercicio anterior se obtuvo mediante muestreo PPT. Calcule una estimación insesgada para el total poblacional de y y calcule el respectivo coeficiente de variación estimado.

12.4 Suponga un diseño de muestreo en dos fases. En la primera fase, se seleccionó una muestra aleatoria simple sin reeemplazo s_a de tamaño $n_a = 160$. En esta fase se estratificó la población en cuatro subgrupos, cada uno de tamaño 40. En la segunda fase, se decidió seleccionar una muestra aleatoria estratificada de 20 elementos en cada estrato y se observó la característica de interés. Los resultados obtenidos se muestran a continuación:

Estrato h	$\bar{y}_{s_h}$	$S^2_{y_{s_h}}$
1	17.05	19945
2	19.75	24179
3	22.40	28359
4	31.25	42829

a. Calcule una estimación insesgada para el total poblacional de y.

b. Obtenga una estimación para la varianza y reporte el respectivo coeficiente de variación estimado.

c. Obtenga una estimación para la varianza y reporte el respectivo coeficiente de variación estimado, suponiendo que la muestra hubiese sido obtenido de un muestreo, en una sola fase, aleatorio estratificado de tamaño $n = 80$.

Capítulo 13

Encuestas multi-propósito

> Si los estadísticos teóricos hacen caso omiso al reto de enfrentar las encuestas multi-propósito, entonces el vacío existente entre ellos y los estadísticos prácticos se hará cada vez más grande. El diseño y análisis de encuestas multivariantes debe ser una de las próximas áreas de mayor investigación.
>
> T. M. F. Smith (1976)

Este capítulo reúne una introducción a la motivación de investigación que hizo uno de los personajes más influyentes en la escena estadística mundial, el profesor Smith (Smith 1976), quien discutió acerca de los fundamentos de la teoría del muestreo, desde sus primeros años hasta las últimas tendencias en cuanto a predicción y estimación en poblaciones finitas. En sus múltiples artículos, este autor afirmó que en el muestreo los problemas univariados (estimación de un parámetro desconocido para una sola característica de interés) se encuentran en unas cuantas ramas de aplicación, limitadas a encuestas de opinión pública, muestreo industrial de aceptación y muestreo en auditorías. Sin embargo, la gran mayoría de encuestas que se realizan alrededor del mundo son de tipo multi-propósito (estimación de varios parámetros desconocidos para varias características de interés). El profesor Smith tuvo en cuenta la limitación que presentan los grandes clásicos del muestreo al no considerar este tipo de estudios ni incluirlos en sus páginas y llamó la atención a los teóricos del muestreo a realizar investigación formal en este tipo de tópicos como lo muestra la motivación al principio del capítulo.

13.1 Introducción

La mayoría de aplicaciones en encuestas por muestreo involucran múltiples variables de estudio. En este breve apartado, se presenta un marco de referencia para la estimación conjunta de los parámetros de interés, bajo algunos diseños de muestreo. Con respecto al diseño de muestreo, en Holmberg (2002*a*) y Holmberg (2002*b*) se ha desarrollado la teoría pertinente para la sección de muestras probabilísticas en encuestas multi-propósito, y con respecto a la estimación multiparamétrica, en Gutiérrez (2009*a*) se propone un sistema general de estimación

basado en resultados clásicos de la teoría de los modelos lineales y del álgebra lineal.

El propósito de un estudio por muestreo está enfocado en obtener información acerca de una población finita particular por medio de la estimación de parámetros poblacionales como medias, totales o proporciones o razones. Sin embargo, la mayoría de encuestas no involucra una sola característica sino varias características de interés. Los libros clásicos de muestreo parecen omitir el hecho de que raras veces se planea una encuesta con el fin de estimar un sólo parámetro y la teoría desarrollada por los investigadores del muestreo está enfocada en la búsqueda de estrategias de muestreo que intentan estimar un parámetro. Existen muchas ventajas desarrolladas en estos tópicos, como se vio en los capítulos anteriores; sin embargo, todas están motivadas bajo el supuesto de que el investigador está interesado en la estimación de un sólo parámetro. Como lo afirma Holmberg (2002a), ≪una encuesta típica en el sector económico involucra varias características de interés y varios parámetros objetivos... con múltiples parámetros de interés y múltiples requerimientos de precisión, el estadístico debería escoger un diseño de muestreo que tenga en cuenta las anteriores características≫.

Una encuesta puede ser divida en dos etapas: la etapa de diseño y la etapa de estimación. El trabajo de Anders Holmberg durante la década pasada está enfocado en la búsqueda de un diseño de muestreo que induzca probabilidades de inclusión desiguales y que sea óptimo en el sentido de que un haya un aumento significativo en la precisión de cada característica de interés. Este capítulo está enfocado en resumir las propuestas de diseño de muestreo y divulgar una posible solución al problema de la estimación multi-paramétrica por medio de un acercamiento matricial para brindar al lector un enfoque exhaustivo de estimación conjunta en muestreo. Aunque los resultados de este capítulo son simples, éstos ofrecen una herramienta poderosa para el planeamiento de estrategias de muestreo en encuestas multi-propósito. En primera instancia, se propondrá el enfoque de estimación en el caso de contar con múltiples características de información auxiliar. Más adelante, se resumirán los resultados de investigación de Holmberg en cuanto al diseño de muestreo de una estrategia que involucra varias características de interés. Por último, el capítulo cierra con un ejemplo numérico que devela el enfoque matricial y sus ventajas en encuestas multi-propósito.

13.2 Estimación de varios parámetros

Suponga que la encuesta involucra el estudio de Q características de interés. Asuma que el k-ésimo elemento ($k \in U$) está asociado a un vector de Q características de interés, $\mathbf{y}_k = (y_{k1}, \ldots, y_{kQ})$ cuyos valores son desconocidos para la población finita. De esta manera, la siguiente matriz será llamada la **matriz de interés**.

$$\mathbf{Y}_U = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1Q} \\ \vdots & \vdots & \ddots & \vdots \\ y_{k1} & y_{k2} & \cdots & y_{kQ} \\ \vdots & \vdots & \ddots & \vdots \\ y_{N1} & y_{N2} & \cdots & y_{NQ} \end{pmatrix} = \begin{pmatrix} \mathbf{y}^1 & \mathbf{y}^2 & \ldots & \mathbf{y}^Q \end{pmatrix} \tag{13.2.1}$$

Note que la entrada y_{kq} se refiere al valor de la q-ésima característica de interés en el k-ésimo elemento, con $k \in U$ y $q = 1, \ldots, Q$. En un contexto de inferencia basada en el diseño de muestreo, $\mathbf{y}^q$ no es considerado como un vector aleatorio, puesto que sus componentes son considerados como parámetros fijos aunque desconocidos. De esta manera, los valores de cada característica de interés no son necesariamente continuos como el ingreso, el peso o la estura sino también discretos como indicadores de subgrupos poblacionales como dominios, estratos o post-estratos. De esta manera, la matriz $\mathbf{Y}_U$ puede ser vista como una matriz de valores mixtos.

El objetivo es estimar los Q componentes del vector de totales definido por la siguiente expresión

$$\mathbf{t} = (t_1, t_2, ..., t_Q)' = \mathbf{Y}_U'\mathbf{1}_N, \tag{13.2.2}$$

donde $\mathbf{1}_N = (1, 1, \ldots, 1)'_{N\times 1}$ y $t_q = \sum_{k\in U} y_{kq}$ es el total poblacional de la q-ésima característica de interés. Cuando la muestra de tamaño n es seleccionada, entonces y_{kq} es observado ($k \in S$) y es posible definir la siguiente matriz

$$\mathbf{Y}_s = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1Q} \\ \vdots & \vdots & \ddots & \vdots \\ y_{k1} & y_{k2} & \cdots & y_{kQ} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nQ} \end{pmatrix}. \tag{13.2.3}$$

Nótese que cuando $s = U$, $\mathbf{Y}_U = \mathbf{Y}_s$. De esta manera, la matriz de probabilidades de inclusión está definida por la siguiente expresión

$$\mathbf{\Pi} = diag(\pi_1, \pi_2, ..., \pi_n), \tag{13.2.4}$$

En este orden de ideas, el estimador de Horvitz-Thompson del vector de totales $\mathbf{t}$ se define como

$$\widehat{\mathbf{t}}_\pi = (\widehat{t}_{1,\pi}, \widehat{t}_{2,\pi}, ..., \widehat{t}_{Q,\pi})' = \mathbf{Y}_s'\mathbf{\Pi}^{-1}\mathbf{1}_n, \tag{13.2.5}$$

con $\mathbf{1}_N = (1, 1, \ldots, 1)'_{n\times 1}$ y $\widehat{t}_{q,\pi} = \sum_{k\in s} y_{kq}/\pi_k$ es el estimador de Horvitz-Thompson de t_q. Es fácil probar que $\widehat{\mathbf{t}}_\pi$ corresponde a un estimador insesgado para $\mathbf{t}$, y su matriz de varianzas está dada por

$$\mathbf{V}(\widehat{\mathbf{t}}_\pi) = E(\widehat{\mathbf{t}}_\pi - \mathbf{t})(\widehat{\mathbf{t}}_\pi - \mathbf{t})'. \tag{13.2.6}$$

Nótese que, si $N \geq q$, entonces $\mathbf{V}(\widehat{\mathbf{t}}_\pi)$ será una matriz simétrica y definida positiva cuyo elemento qq' es

$$\sum_{k\in U}\sum_{l\in U} \Delta_{kl}\frac{y_{kq}}{\pi_k}\frac{y_{lq'}}{\pi_l}, \tag{13.2.7}$$

con $\Delta_{kl} = \pi_{kl} - \pi_k\pi_l$. Si $s \neq U$ es imposible calcular el valor de la anterior expresión. Sin embargo, si $n \geq q$, la varianza puede ser estimada mediante una matriz simétrica y definida positiva $\widehat{\mathbf{V}}(\widehat{\mathbf{t}}_\pi)$ cuyo elemento qq' es

$$\sum_{k\in S}\sum_{l\in s}\frac{\Delta_{kl}}{\pi_{kl}}\frac{y_{kq}}{\pi_k}\frac{y_{lq'}}{\pi_l}. \tag{13.2.8}$$

En algunos casos, el requerimiento de la encuesta es la estimación del vector de medias poblacionales dado por

$$\bar{\mathbf{y}} = \frac{1}{N}\mathbf{t}. \tag{13.2.9}$$

Por lo tanto, un estimador insesgado para $\bar{\mathbf{y}}$ es

$$\bar{\mathbf{y}}_\pi = \frac{1}{N}\widehat{\mathbf{t}}_\pi, \tag{13.2.10}$$

cuya matriz de varianzas será estimada insesgadamente por $\frac{1}{N^2}\widehat{\mathbf{V}}(\widehat{\mathbf{t}}_\pi)$. si el tamaño poblacional es desconocido, entonces puede ser estimada insesgadamente usando los principio del estimador de Horvitz-Thompson, tal que

$$\widehat{N}_\pi = \mathbf{1}'_n\mathbf{\Pi}^{-1}\mathbf{1}_n. \tag{13.2.11}$$

Note que la eficiencia computacional podría aumentarse con la incorporación de este enfoque matricial puesto que la estimación de varios parámetros de interés se realiza mediante una sola operación algebraica.

13.3 Algunos diseños de muestreo

En esta sección se introducen algunos ejemplos de estimación de varios parámetros de interés bajo los diseños de muestreo más comunes en la teoría.

Resultado 13.3.1. *Bajo el diseño de muestreo Bernoulli, el vector de totales* $\mathbf{t}$ *es estimado insesgadamente por*

$$\widehat{\mathbf{t}}_\pi = \frac{1}{\pi}\mathbf{Y}'_s\mathbf{1}_n \tag{13.3.1}$$

y su matriz de varianzas es estimada insesgadamente por

$$\widehat{\mathbf{V}}(\widehat{\mathbf{t}}_\pi) = \frac{1}{\pi}\left(\frac{1}{\pi}-1\right)\mathbf{Y}_s\mathbf{Y}'_s. \tag{13.3.2}$$

Resultado 13.3.2. *Aunque el diseño de muestreo aleatorio simple sin reemplazo no es el más utilizado en la práctica, sí es utilizado en las últimas etapas de muestreo en diseños complejos. Bajo este diseño de muestreo* $\mathbf{t}$ *es estimado insesgadamente por*

$$\widehat{\mathbf{t}}_\pi = \frac{N}{n}\mathbf{Y}'_s\mathbf{1}_n. \tag{13.3.3}$$

y su matriz de covarianzas es estimada insesgadamente por

$$\widehat{\mathbf{V}}(\widehat{\mathbf{t}}_\pi) = \frac{N^2}{n}\left(1-\frac{n}{N}\right)\mathbf{S}_y, \tag{13.3.4}$$

con $\mathbf{S}_y$, la matriz de covarianzas de las característica de interés calculada con las observaciones recolectadas en la muestra seleccionada. Por otro lado, $\bar{\mathbf{y}}$ es estimada insesgadamente por

$$\bar{\mathbf{y}}_\pi = \frac{1}{N}\widehat{\mathbf{t}}_\pi = \frac{1}{n}\mathbf{Y}_s'\mathbf{1}_n. \tag{13.3.5}$$

y su matriz de covarianzas es estimada insesgadamente mediante la siguiente expresión

$$\widehat{\mathbf{V}}(\bar{\mathbf{y}}_\pi) = \frac{1}{N^2}\widehat{\mathbf{V}}(\widehat{\mathbf{t}}_\pi). \tag{13.3.6}$$

13.3.1 Estimación en dominios

Si los requerimientos de la encuesta están relacionados con la estimación del tamaño absoluto de un dominio o del total de alguna o varias características de interés en tal dominio, entonces se propone la siguiente construcción metodológica. Suponga que la población está particionada en D dominios tales que $U = U_1, \ldots, U_d, \ldots, U_D$. Entonces, se define la **matriz indicadora de dominios** como

$$\mathbf{Z} = \begin{pmatrix} z_{11} & \ldots & z_{1d} & \ldots & z_{1D} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ z_{k1} & \ldots & z_{kd} & \ldots & z_{kD} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ z_{n1} & \ldots & z_{nd} & \ldots & z_{nD} \end{pmatrix} \tag{13.3.7}$$

donde el elemento

$$z_{kd} = \begin{cases} 1 & \text{si } k \in U_d, \text{ y} \\ 0 & \text{en otro caso} \end{cases} \tag{13.3.8}$$

El vector de tamaños absolutos del dominio d está dado por

$$\mathbf{N}_d = (N_1, N_2, ..., N_D)' \tag{13.3.9}$$

donde

$$N_d = \sum_{k \in U} z_{kd}. \tag{13.3.10}$$

$\mathbf{N}_d$ es estimado insesgadamente por el estimador de Horvitz-Thompson de la siguiente manera

$$\widehat{\mathbf{N}}_d = (\widehat{N}_1, \widehat{N}_2, ..., \widehat{N}_D)' = \mathbf{Z}'\mathbf{\Pi}^{-1}\mathbf{1}_n, \tag{13.3.11}$$

su matriz de varianzas es estimada insesgadamente por $\widehat{\mathbf{V}}(\widehat{\mathbf{N}}_d)$, la cual está definida análogamente a (13.2.6).

En muchas ocasiones se requiere de la estimación de los totales de características de interés sobre todos los dominios. De esta forma, el total de la q-ésima variable sobre todos los D dominios de interés está dado por

$$\mathbf{t}_{dq} = (t_{1q}, t_{2q}, ..., t_{Dq})' \tag{13.3.12}$$

y una forma de estimarlo está dada por la siguiente expresión

$$\widehat{\mathbf{t}}_{dq\pi} = (\widehat{t}_{1q\pi}, \widehat{t}_{2q\pi}, ..., \widehat{t}_{Dq\pi})' = (\mathbf{y}^q \mathbf{1}_D \odot \mathbf{Z})' \mathbf{\Pi}^{-1} \mathbf{1}_n \tag{13.3.13}$$

En donde, $\mathbf{y}^q$ denota la q-ésima columna de la matriz $\mathbf{Y}_s$, $\mathbf{1}_D = (1, \ldots, 1)'_{D\times 1}$ y $\odot$ denota el producto matricial de Hadamard.

Resultado 13.3.3. *Bajo el diseño de muestreo aleatorio simple sin reemplazo, el estimador de Horvitz-Thompson para el vector de tamaños absolutos de los dominios y para el total de la q-ésima característica de interés en todos los D dominios están dados por*

$$\widehat{\mathbf{N}}_d = (N/n)\mathbf{Z}'\mathbf{1}_n, \tag{13.3.14}$$

$$\widehat{\mathbf{t}}_{dq\pi} = (N/n)(\mathbf{y}^q \mathbf{1}_D \odot \mathbf{Z})\mathbf{1}_n. \tag{13.3.15}$$

respectivamente.

13.3.2 Estimación en diseños estratificados

Para diseños estratificados se tiene el siguiente marco de referencia. La población finita U se divide en H grupos o estratos mutuamente excluyentes $U_1, \ldots, U_h \ldots, U_H$. Note que antes de la recolección de los datos, se conoce la membresía de cada elemento a cada estrato. De esta manera, se selecciona una muestra aleatoria en todos y cada uno de los H estratos existentes en la población finita. Es necesario realizar un ordenamiento matricial en las matrices para obtener estimaciones usando los principios del estimador de Horvitz-Thompson estimator. Por lo tanto, la matriz $\mathbf{Y}_s$ se particiona en H bloques de la siguiente manera

$$\mathbf{Y}_s = \begin{pmatrix} \mathbf{Y}_1 \\ \vdots \\ \mathbf{Y}_h \\ \vdots \\ \mathbf{Y}_H \end{pmatrix}, \tag{13.3.16}$$

donde $\mathbf{Y}_h$ es una submatriz que contiene los valores de cada característica de interés para los elementos que pertenecen al h-ésimo estrato, con $h = 1, \ldots, H$. Note que $\mathbf{Y}_s \in \Re^{Hn\times Q}$ y $\mathbf{Y}_h \in \Re^{n_h\times Q}$. Definido $\mathbf{n} = (n_1, \ldots, n_H)'$, entonces $n = \mathbf{n}'\mathbf{1}_H = n_1 + \cdots + n_H$.

Como de costumbre, el objetivo es la estimación de los Q componentes del vector de totales en el h-ésimo estrato dado por

$$\mathbf{t}_h = (t_{1h}, t_{2h}, ..., t_{Qh})' = \mathbf{Y}'_h \mathbf{1}_{N_h}, \tag{13.3.17}$$

Donde N_h es el tamaño del h-ésimo estrato. El total poblacional puede ser escrito como

$$\mathbf{t} = (t_1, t_2, ..., t_Q)' = \sum_{h=1}^{H} \mathbf{t}_h, \tag{13.3.18}$$

donde $\mathbf{t}_h$ es estimado insesgadamente por la siguiente expresión

$$\widehat{\mathbf{t}}_{h\pi} = (\widehat{t}_{1h\pi}, \widehat{t}_{2h\pi}, ..., \widehat{t}_{Qh\pi})' = \mathbf{Y}'_h \mathbf{1}_{n_h}, \tag{13.3.19}$$

con n_h el tamaño de la muestra en el h-ésimo estrato. Por supuesto, se asume independencia sobre el diseño de muestreo implementado en cada estrato. De esta forma el total poblacional está dado por

$$\widehat{\mathbf{t}}_\pi = (\widehat{t}_{1\pi}, \widehat{t}_{2\pi}, ..., \widehat{t}_{Q\pi})' = \sum_{h=1}^{H} \widehat{\mathbf{t}}_h, \tag{13.3.20}$$

y su matriz de varianzas puede ser escrita como

$$\mathbf{V}_{ST}(\widehat{\mathbf{t}}_\pi) = \sum_{h=1}^{H} \mathbf{V}_h(\widehat{\mathbf{t}}_\pi) \tag{13.3.21}$$

la cual es estimada insesgadamente por

$$\widehat{\mathbf{V}}_{ST}(\widehat{\mathbf{t}}_\pi) = \sum_{h=1}^{H} \widehat{\mathbf{V}}_h(\widehat{\mathbf{t}}_\pi). \tag{13.3.22}$$

Resultado 13.3.4. *Bajo el diseño de muestreo aleatorio estratificado, el estimador de Horvitz-Thompson para el total poblacional es*

$$\widehat{\mathbf{t}}_\pi = \sum_{h=1}^{H} \frac{N_h}{n_h} \mathbf{Y}_h' \mathbf{1}_{n_h}, \tag{13.3.23}$$

y su matriz de covarianzas es estimada insesgadamente por

$$\widehat{\mathbf{V}}_{STSI}(\widehat{\mathbf{t}}_\pi) = \sum_{h=1}^{H} \frac{N_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right) \mathbf{S}_{yh}, \tag{13.3.24}$$

con $\mathbf{S}_{yh}$, la matriz de varianzas de las características de interés en la muestra perteneciente al h-ésimo estrato.

13.4 Información auxiliar

Asuma que el k-ésimo elemento ($k \in U$) está asociado con un vector de P características de información auxiliar, contenidas en un vector $\mathbf{x}_k$. Los valores de este vector $\mathbf{x}_k = (x_{k1}, \ldots, x_{kP})$ se suponen conocidos para la población finita. De esta manera se tiene la siguiente matriz

$$\mathbf{X}_U = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1P} \\ \vdots & \vdots & \ddots & \vdots \\ x_{k1} & x_{k2} & \ldots & x_{kP} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \ldots & x_{NP} \end{pmatrix} = \begin{pmatrix} \mathbf{x}^1 & \mathbf{x}^2 & \ldots & \mathbf{x}^P \end{pmatrix} \tag{13.4.1}$$

que será llamada la **matriz de información auxiliar**.

13.4.1 Algunos relaciones

Es posible asumir que existe una relación lineal explícita entre cada uno de los componentes de las características de interés y las características de información auxiliar mediante un modelo de superpoblación ξ_q, $q = 1, \ldots, Q$, tal que

$$\underset{(N\times 1)}{\mathbf{Y}^q} = \underset{(N\times P)}{\mathbf{X}}\underset{(P\times 1)}{\boldsymbol{\beta}^q} + \underset{(N\times 1)}{\boldsymbol{\varepsilon}^q}.$$

El modelo ξ_q tiene las siguientes propiedades:

$$\begin{aligned} E_{\xi_q}(\boldsymbol{\varepsilon}^q) &= \mathbf{0} \\ Var_{\xi_q}(\boldsymbol{\varepsilon}^q) &= \boldsymbol{\Sigma}_q. \end{aligned} \tag{13.4.2}$$

Σ_q establece la estructura de varianza del vector $\boldsymbol{\varepsilon}^q$. Nótese que las anteriores relaciones pueden reescribirse mediante un modelo conjunto ξ tal que

$$\underset{(N\times Q)}{\mathbf{Y}} = \underset{(N\times P)}{\mathbf{X}}\underset{(P\times Q)}{\boldsymbol{\beta}} + \underset{(N\times Q)}{\boldsymbol{\varepsilon}}.$$

Este enfoque sugiere que $\mathbf{Y}$, $\mathbf{X}$ y $\boldsymbol{\varepsilon}$ son matrices aleatorias (Gupta & Nagar 1999) definidas en el modelo de superpoblación ξ, para el cual $\mathbf{Y}_U$ y $\mathbf{X}_U$ se suponen meras realizaciones de las anteriores matrices aleatorias. Más precisamente, el modelo ξ tiene las siguientes características:

$$\begin{aligned} E_{\xi}(\boldsymbol{\varepsilon}) &= \underset{(N\times Q)}{\mathbf{0}} \\ Var_{\xi}(\vec{\boldsymbol{\varepsilon}}) &= \underset{(NQ\times NQ)}{\boldsymbol{\Sigma}} = diag(\Sigma_1, \Sigma_2, \ldots, \Sigma_Q) \end{aligned} \tag{13.4.3}$$

Note que el subíndice ξ se refiere a la esperanza bajo la estructura particular que ese modelo de superpoblación induce. En situaciones prácticas, es común asumir $\boldsymbol{\Sigma}_q = \sigma_q^2 diag(c_{1q}, \ldots, c_{Nq})$, donde $c_{kq} = f_q(x_{k1}, \ldots, x_{kP})$ y f_q es una función de valor real.

El problema de estimar el vector de parámetros $\boldsymbol{\beta}$ se considera brevemente. Sea $D(\mathbf{X})$ una medida de dispersión invariante ante traslaciones tal que $D(\mathbf{X} + \mathbf{K}) = D(\mathbf{X})$, con $\mathbf{K}$ una matriz de constantes. Entonces al estimación de $\boldsymbol{\beta}$ corresponderá a aquel vector que minimize la anterior medida de dispersión. Particularmente, $D(\cdot)$ podría estar dada por la varianza total multivariante definida como

$$traza(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}). \tag{13.4.4}$$

Con la anterior elección y recurriendo al método de mínimos cuadrados, (13.4.4) es minimizada por la siguiente expresión

$$\mathbf{B} = (\mathbf{B}_1, \mathbf{B}_2, \ldots, \mathbf{B}_Q), \tag{13.4.5}$$

donde

$$\mathbf{B}_q = (\mathbf{X}_U'\Sigma_q^{-1}\mathbf{X}_U)^{-1}(\mathbf{X}_U'\Sigma_q^{-1}\mathbf{Y}_U). \tag{13.4.6}$$

Nótese que para poder calcular esta estimación, se deben conocer todos los valores poblacionales tanto de la matriz de características de interés como de las características de información auxiliar.

13.4.2 Información tradicional

En aplicaciones reales sólo se selecciona una muestra y no es posible calcular $\mathbf{B}$. Por lo tanto, este valor debe ser estimado recurriendo a la información disponible en la muestra aleatoria seleccionada o realizada. Puede ser demostrado que la siguiente expresión corresponde a un estimador asintóticamente insesgado para $\mathbf{B}$

$$\widehat{\mathbf{B}} = (\widehat{\mathbf{B}_1}, \widehat{\mathbf{B}_2}, \ldots, \widehat{\mathbf{B}_Q}), \tag{13.4.7}$$

donde

$$\widehat{\mathbf{B}_q} = (\mathbf{X}_s' \mathbf{A}_q^{-1} \mathbf{X}_s)^{-1} (\mathbf{X}_s' \mathbf{A}_q^{-1} \mathbf{Y}_s), \tag{13.4.8}$$

$q = 1, \ldots, Q$, $\mathbf{X_s}$ similarmente definido como en (13.2.3) y

$$\mathbf{A}_q = \mathbf{\Pi}^{1/2} \Sigma_q \mathbf{\Pi}^{1/2}. \tag{13.4.9}$$

Así, el **estimador múltiple de regresión general** para el vector de totales poblacionales se define como

$$\widehat{\mathbf{t}}_{Mgreg} = \widehat{\mathbf{t}}_{\mathbf{y}\pi} + \widehat{\mathbf{B}}' (\mathbf{t}_{\mathbf{x}} - \widehat{\mathbf{t}}_{\mathbf{x}\pi}), \tag{13.4.10}$$

con, $\widehat{\mathbf{t}}_{\mathbf{y}\pi}$, $\widehat{\mathbf{t}}_{\mathbf{x}\pi}$ los estimadores de Horvitz-Thompson de $\mathbf{t_y}$ y $\mathbf{t_x}$, respectivamente. Nótese que $\widehat{\mathbf{B}}_q$ también puede ser escrito como

$$\widehat{\mathbf{B}}_q = (\mathbf{X}_s' \mathbf{D}_\lambda \mathbf{X}_s)^{-1} \mathbf{X}_s \mathbf{D}_\lambda \mathbf{Y}_s \tag{13.4.11}$$

$$= \left(\sum_{k \in s} \mathbf{x}_k \lambda_k^q \mathbf{x}_k' \right)^{-1} \left(\sum_{k \in s} \mathbf{x}_k \lambda_k^q \mathbf{y}_k' \right) \tag{13.4.12}$$

donde $\mathbf{D}_\lambda = diag(\lambda_1^q, \ldots, \lambda_n^q)$ y λ_k^q son funciones de valor real de las probabilidades de inclusión y de la información auxiliar. Note también que el modelo ξ sirve como un vehículo para encontrar un estimador de regresión general apropiado. Una vez que éste se encuentra o se define, el modelo no será útil para ningún otro propósito de muestreo. Las propiedades del estimador múltiple de regresión general (esperanza y varianza) también se definen desde una perspectiva de inferencia basada en el diseño de muestreo.

Algunos casos particulares

Los siguientes escenarios se enuncian bajo un marco de referencia general que resultan ser casos especiales del estimador múltiple de regresión general; en la mayoría de los casos su particularidad está inducida por la escogencia de los valores de λ_k.

- Si $P = 1$, $\mathbf{x}_k = x_k$, y $\lambda_k^q = (\pi_k x_k)^{-1}$, entonces se tiene el estimador de razón para cada característica de interés.
- Si $P = 2$, $\mathbf{x}_k = (1, x_k)'$, y $\lambda_k^q = (\pi_k)^{-1}$, entonces se tiene el estimador de regresión clásico.

- Si $P = M$(number of post-strata), $\mathbf{x}_k = \delta_k = (0, \ldots 0, 1, 0, \ldots, 0)'$, y $\lambda_k^q = (\pi_k)^{-1}$, donde δ_k representa M variables indicadoras (cada indicadora representa la membresía del elemento poblacional al post-estrato en cuestión), entonces tenemos el estimador de post-estratificación.

Nótese que el estimador múltiple de regresión general puede también escribirse de la siguiente manera

$$\widehat{\mathbf{t}}_{Mgreg} = (\mathbf{W}' \odot \mathbf{Y}_s')\mathbf{1}_n, \tag{13.4.13}$$

donde

$$\mathbf{W} = \begin{pmatrix} w_1^1 & w_1^2 & \ldots & w_1^Q \\ \vdots & \vdots & \ddots & \vdots \\ w_k^1 & w_k^2 & \ldots & w_k^Q \\ \vdots & \vdots & \ddots & \vdots \\ w_n^1 & w_n^2 & \ldots & w_n^Q \end{pmatrix} = \begin{pmatrix} \mathbf{w}^1 & \mathbf{w}^2 & \ldots & \mathbf{w}^Q \end{pmatrix}. \tag{13.4.14}$$

Se tiene que $\mathbf{w}^q = (w_1^q, \ldots, w_k^q, \ldots, w_n^q)'$ es un vector de pesos o ponderaciones tales que

$$w_k^q = \frac{1}{\pi_k}\left(1 + \lambda_k^q \mathbf{x}_k' \left(\sum_{k \in s} \mathbf{x}_k \lambda_k^q \mathbf{x}_k'\right)^{-1} (\mathbf{t_x} - \widehat{\mathbf{t}}_{\mathbf{x}\pi})\right). \tag{13.4.15}$$

A estos pesos, como se estudió en capítulos anteriores, se le conocen con el nombre de **ponderaciones de calibración** y ellos reproducen con exactitud el vector de totales $\mathbf{t_x}$ cuando son aplicados a la información auxiliar disponible. Entonces, $\mathbf{W}$ es llamada **matriz de calibración**. No es difícil mostrar que la siguiente relación

$$\sum_{k \in S} w_k^q \mathbf{x}_k = \mathbf{X}_s' \mathbf{w}^q = \mathbf{t_x}, \tag{13.4.16}$$

se satisface para cada $q = 1, \ldots, Q$. Es interesante observar que $\mathbf{t_x}$ resulta calibrado bajo diferentes escogencias de los pesos w^q. Por otra parte, note que

$$\mathbf{w}^q = \mathbf{\Pi}^{-1}\mathbf{1}_n + \mathbf{A}_q \mathbf{X}_s \left(\mathbf{X}_s' \mathbf{A}_q \mathbf{X}_s\right)^{-1} (\mathbf{t_x} - \widehat{\mathbf{t}}_{\mathbf{x}\pi}) \tag{13.4.17}$$

Cuando se trata de estimación post-estratificada se debe recurrir al uso de una inversa generalizada, acudiendo a la propiedad de que el estimador múltiple de general de regresión es invariante ante cualquier inversa.

13.4.3 Información auxiliar conjunta

El método de mínimos cuadrados no es el único camino para obtener un estimador múltiple de regresión general. En esta sección, se supone la existencia de una matriz de información conjunta cuya estructura algebraica está definida por la siguiente expresión

$$\mathbf{V} = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1Q} & x_{11} & x_{12} & \cdots & x_{1P} \\ y_{21} & y_{22} & \cdots & y_{2Q} & x_{21} & x_{22} & \cdots & x_{2P} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nQ} & x_{n1} & x_{n2} & \cdots & x_{nP} \end{pmatrix}. \tag{13.4.18}$$

El estimador del vector de totales de las características de interés y de las características de información auxiliar está dado por $\widehat{\mathbf{t}}_{\mathbf{v}\pi}$, el cual está definido como

$$\widehat{\mathbf{t}}_{\mathbf{v}\pi} = \mathbf{V}'\mathbf{\Pi}^{-1}\mathbf{1}_n. \tag{13.4.19}$$

De esta forma, suponga que $\widehat{\mathbf{t}}_{\mathbf{v}\pi}$ sigue una distribución normal multivariante con media

$$E\left(\widehat{\mathbf{t}}_{\mathbf{v}\pi}\right) = (\mathbf{t}'_{\mathbf{Y}\pi}, \mathbf{t}'_{\mathbf{X}\pi})' = \mathbf{t}_{\mathbf{v}},$$

y matriz de varianzas definida como

$$V\left(\widehat{\mathbf{t}}_{\mathbf{v}\pi}\right) = \begin{pmatrix} V(\widehat{\mathbf{t}}_{\mathbf{y}\pi}) & C(\widehat{\mathbf{t}}_{\mathbf{y}\pi}, \widehat{\mathbf{t}}_{\mathbf{x}\pi}) \\ C(\widehat{\mathbf{t}}_{\mathbf{y}\pi}, \widehat{\mathbf{t}}_{\mathbf{x}\pi}) & V(\widehat{\mathbf{t}}_{\mathbf{x}\pi}) & , \end{pmatrix},$$

donde $V(\widehat{\mathbf{t}}_{\mathbf{y}\pi})$ se considera una matriz simétrica tal que el j-ésimo elemento de su diagonal está dado por la varianza de $\widehat{t}_{y_j\pi}$

$$V(\widehat{t}_{y_j\pi}) = \sum\sum_U \Delta_{kl}\frac{y_{jk}}{\pi_k}\frac{y_{jl}}{\pi_l},$$

y el elemento ij, por afuera de su diagonal, está dado por la covarianza de $\widehat{t}_{y_i\pi}$ y $\widehat{t}_{y_j\pi}$,

$$C(\widehat{t}_{y_i\pi}, \widehat{t}_{y_j\pi}) = \sum\sum_U \Delta_{kl}\frac{y_{ik}}{\pi_k}\frac{y_{jl}}{\pi_l}.$$

$V(\widehat{\mathbf{t}}_{\mathbf{x}\pi})$ se define de forma análoga, y $C(\widehat{\mathbf{t}}_{\mathbf{y}\pi}, \widehat{\mathbf{t}}_{\mathbf{x}\pi})$, no necesariamente simétrica, es tal que el elemento ij está dado por la covarianza de $\widehat{t}_{y_i\pi}$ and $\widehat{t}_{x_j\pi}$

$$C(\widehat{t}_{y_i}\pi, \widehat{t}_{x_j}\pi) = \sum\sum_U \Delta_{kl}\frac{y_{ik}}{\pi_k}\frac{x_{jl}}{\pi_l}.$$

Siguiendo los resultados de inferencia multivariante para poblaciones con distribución normal, la distribución condicional de $\widehat{\mathbf{t}}_{\mathbf{Y}\pi}$ dado $\widehat{\mathbf{t}}_{\mathbf{X}\pi}$ sigue también una distribución normal multivariante con media condicional dada por

$$E(\widehat{\mathbf{t}}_{\mathbf{y}\pi}|\widehat{\mathbf{t}}_{\mathbf{x}\pi}) = \mathbf{t}_{\mathbf{y}\pi} + C(\widehat{\mathbf{t}}_{\mathbf{y}\pi}, \widehat{\mathbf{t}}_{\mathbf{x}\pi})(V(\widehat{\mathbf{t}}_{\mathbf{x}\pi}))^{-1}(\mathbf{t}_{\mathbf{x}} - \widehat{\mathbf{t}}_{\mathbf{x}\pi}), \tag{13.4.20}$$

Y varianza condicional dada por

$$V(\widehat{\mathbf{t}}_{\mathbf{y}\pi}|\widehat{\mathbf{t}}_{\mathbf{x}\pi}) = V(\widehat{\mathbf{t}}_{\mathbf{y}\pi}) - C(\widehat{\mathbf{t}}_{\mathbf{y}\pi}, \widehat{\mathbf{t}}_{\mathbf{x}\pi})(V(\widehat{\mathbf{t}}_{\mathbf{x}\pi}))^{-1}C(\widehat{\mathbf{t}}_{\mathbf{x}\pi}, \widehat{\mathbf{t}}_{\mathbf{y}\pi}). \tag{13.4.21}$$

Note que (13.4.20) y (13.4.21) son estimados insesgadamente por

$$\widehat{\mathbf{t}}_{\mathbf{y}} = \widehat{\mathbf{t}}_{\mathbf{y}\pi} + \widehat{C}(\widehat{\mathbf{t}}_{\mathbf{y}\pi}, \widehat{\mathbf{t}}_{\mathbf{x}\pi})(\widehat{V}(\widehat{\mathbf{t}}_{\mathbf{x}\pi}))^{-1}(\mathbf{t}_{\mathbf{x}} - \widehat{\mathbf{t}}_{\mathbf{x}\pi}) \tag{13.4.22}$$

$$= \widehat{\mathbf{t}}_{\mathbf{y}\pi} + \widehat{\mathbf{B}}(\mathbf{t}_{\mathbf{x}} - \widehat{\mathbf{t}}_{\mathbf{x}\pi}) \tag{13.4.23}$$

y,

$$\widehat{V}(\widehat{\mathbf{t}}_{\mathbf{y}}) = V(\widehat{\mathbf{t}}_{\mathbf{Y}\pi}) - C(\widehat{\mathbf{t}}_{\mathbf{Y}\pi}, \widehat{\mathbf{t}}_{\mathbf{X}\pi})(V(\widehat{\mathbf{t}}_{\mathbf{X}\pi}))^{-1}C(\widehat{\mathbf{t}}_{\mathbf{X}\pi}, \widehat{\mathbf{t}}_{\mathbf{Y}\pi}), \tag{13.4.24}$$

respectivamente. Por otra parte, observe que (13.4.22) luce como el estimador múltiple de regresión general. Sin embargo, su pendiente, $\widehat{\mathbf{B}}$, sería diferente: mientras la pendiente del estimador de regresión general está dada por el método de mínimos cuadrados, la pendiente de éste último corresponde, según los resultados de la inferencia estadística multivariante, a un conjunto de regresiones múltiples de $\mathbf{X}$ sobre $\mathbf{Y}$. Este estimador del vector de totales de la característica de interés debería ser llamado **estimador óptimo de regresión general** y ha sido estudiado por Cassady & Valliant (1993) en el contexto de la inferencia basada en modelos poblacionales para la estimación del total de una sola característica de interés.

13.5 Diseños de muestreo óptimos

En esta sección se aborda el problema de la escogencia de la muestra bajo un criterio unificado que contemple el comportamiento estructural de cada una de las características de interés. Es decir, en la etapa de diseño de una encuesta multipropósito se debe escoger un diseño de muestreo integral y para esto el enfoque de Holmbersg será considerado. De esta manera, se puede asumir que en la etapa de planeación de la estrategia, es posible contar con la participación de características de información auxiliar y con esto es posible asumir algunas posturas acerca de la validez de las relaciones estadísticas entre las características de interés y las variables de información auxiliar.

13.5.1 Diseño de muestreo de Holmberg

Suponga que las características de interés involucradas en la encuestas tienen todas la misma importancia[1]. Bajo este enfoque se presenta a continuación un breve resumen del diseño de Holmberg utilizado en encuestas multi-propósito:

1. Para cada una de las características de interés, el estadístico, el investigador o el usuario final debe proponer un diseño de muestreo, $p_q(\cdot)$ $(q = 1, \ldots, Q)$, que sea óptimo y tal que el tamaño esperado de muestra sea $E(n(S)) = n_q$. Por supuesto, note que cada uno de los Q diseños de muestreo pueden ser diferentes; aún más, los tamaños de muestra, en cada diseño propuesto, no necesariamente deben ser equivalentes. Recuerde que el enfoque tradicional, que no se preocupa por la inclusión de varias características de interés, el estadístico debe proponer un sólo diseño de muestreo, el cual se supone que es óptimo para todos los parámetros que se deben estimar.

2. Cada uno de los diseños de muestreo $p_q(\cdot)$ induce un vector de probabilidades de inclusión de tamaño N para cada una de los elementos pertenecientes a la población finita. Estas probabilidades de inclusión deben tomar la siguiente forma (Holmberg 2002*b*, eq. 6)

$$\pi_{qk} = n_q \frac{\sigma_{qk}}{\sum_{k \in S} \sigma_{qk}}, \tag{13.5.1}$$

[1] Por supuesto, que es posible asumir variantes ante este supuesto y pueden ser consultadas en Holmberg (2002*b*). sin embargo, en este capítulo se asumirá que la encuesta contempla igual importancia para todas las características de interés.

con σ_{qk} medidas de tamaño (usualmente, aunque no necesariamente, vinculadas a un modelo de regresión lineal). La característica de «diseño de muestreo óptimo» se obtiene si $\pi_{qk} \propto \sigma_{qk}$. Note que si el diseño de muestreo óptimo para la q-ésima característica de interés es un diseño de muestreo aleatorio simple sin reemplazo, entonces $\sigma_{qk} = 1$ para todo $k \in U$. Por otra parte, con la escogencia de $\sigma^2_{qk} = \sigma^2_q x^{\gamma_q}_{qk}$, donde σ^2_q es una constante y x_{qk} corresponde al valor del k-ésimo elemento para alguna variable auxiliar, o una función de muchas variables de información auxiliar, entonces el diseño de muestreo óptimo debe ser proporcional al tamaño de σ_{qk} (πPS). Es decir, $\pi_{qk} \propto x^{\gamma_q/2}_{qk}$.

3. Basado en el criterio de mínima pérdida de eficiencia relativa general (ANOREL, por sus siglas en inglés), el tamaño de muestra óptimo para la encuesta multipropósito estará dado por

$$n^* \geq \frac{(\sum_{k\in U}\sqrt{a_{qk}})^2}{(1+c)Q + \sum_{k\in U} a_{qk}}, \qquad (13.5.2)$$

donde

$$a_{qk} = \sum_{q=1}^{Q} \frac{\sigma^2_{qk}}{\sum_{k\in U}\left(\frac{1}{\pi_{qk}} - 1\right)\sigma^2_{qk}}, \qquad (13.5.3)$$

y c es el máximo error permitido, bajo el criterio ANOREL, en una escala de cero hasta uno. Nótese que en la práctica, σ^2_{qk} es desconocido y debe ser escrito como una función de las variables de información auxiliar. Holmberg (2002*b*) afirma que el conocimiento subjetivo, la experiencia, o fuentes externas pueden ser usadas para obtener acercamientos al valor exacto de esta cantidad.

4. Una vez que el tamaño de la muestra ha sido calculado, se debe crear un sólo vector de probabilidades de inclusión que sea óptimo para todas las características de interés. Este vector es inducido por el diseño de muestreo de Holmberg, el cual minimiza la pérdida de eficiencia relativa general, está dado por la siguiente expresión

$$\pi_{(opt)k} = \frac{n^*\sqrt{a_{qk}}}{\sum_{k\in U}\sqrt{a_{qk}}} \qquad (13.5.4)$$

5. En la mayoría de los casos, el vector de probabilidades de inclusión resultante, $\boldsymbol{\pi}_{(opt)} = (\pi_{(opt)1}, \ldots, \pi_{(opt)N})'$, es un vector de probabilidades de inclusión desiguales. En esta situación, se debe usar un esquema de selección de muestras πPT

13.5.2 Un ejemplo numérico

En esta sección, se considera un ejemplo del enfoque multi-propósito. En la etapa de diseño, se escoge un diseño de muestreo óptimo por medio del enfoque de Holmberg (Holmberg 2002*b*) y en la etapa de estimación se implementa el enfoque matricial (Gutiérrez 2009*a*). Ambas etapas se realizan por medio del software

computacional R. Particularmente, se introduce el paquete `sampling` para la selección de muestras y la estimación en varios dominios de interés.

Para este propósito, se considera una población real (la población de municipalidades suizas MU281 disponible en el apéndice B de Särndal, Swensson & Wretman (1992)). De esta forma, es posible planear una encuesta multi-propósito en donde las características de interés y los dominios de interés son provistos de antemano y en donde es posible tener cierta clase de creencias acerca del comportamiento estructural de la población y acerca de la relación entre las características de interés y las de información auxiliar. Nótese que no se quiere presentar un diseño de muestreo perfecto, pero más bien uno que ilustre el desarrollo práctico de la teoría en una encuesta multi-propósito. Las características de interés son:

y_1 = P85 (Población en 1985)
y_2 = RMT85 (Impuestos devengados por los municipios en 1985)
y_3 = REV84 (Valores de bienes raíces en 1984)

Las características de información auxiliar son:

x_1 = P75 (Población en 1975)
x_2 = S82 (Número de curules en el consejo de los municipios en 1982)

Para la estimación por dominios se utiliza la siguiente variable:

z = REG (indicador de región geográfica)

Se utilizó el siguiente código computacional para especificar las características de la encuesta.

```
> library(sampling)
> data(MU284)
> MU281 <- MU284[MU284$RMT85 <= 3000,]
> attach(MU281)

> Y1 <- P85
> Y2 <- RMT85
> Y3 <- REV84
> X1 <- P75
> X2 <- S82
> Z  <- REG
```

Para tener algún grado de certeza acerca de las bondades de la estimación se tienen a la mano los totales de las características de interés y de información auxiliar.

```
> Ty <- c(sum(Y1),sum(Y2),sum(Y3))
> Tx <- c(N,sum(X1),sum(X2))

> Ty
[1]   7033  53151 757246
> Tx
[1]   281   6818 13257
```

Ahora, suponiendo que la importancia de las tres características de interés es la misma, entonces a continuación se describe el enfoque de Holmberg para este caso particular:

1. En la población MU281, el tamaño poblacional es $N = 281$. Suponga que el estadístico considera que para cada una de las tres características de interés el tamaño de la muestra debe ser igual a 100.

```
> N <- 281; n <- 100
```

2. Asuma que, mediante conocimiento de fuentes externas, el estadístico asume que los mejores diseños de muestreo, en el sentido óptimo, son: Para y_1, un diseño de muestreo πPT con $\pi_{1k} \propto x_{1k}^{0.7}$, para y_2, un diseño de muestreo πPT con $\pi_{2k} \propto x_{1k}$ y por último para y_3, un diseño de muestreo aleatorio simple.

```
> sigy1 <- sqrt(X1^(1.4))
> sigy2 <- sqrt(X1^(2))
> sigy3 <- rep(1,N)

> pik1 <- n*sigy1/(sum(sigy1))
> pik2 <- n*sigy2/(sum(sigy2))
> pik3 <- n*sigy3/(sum(sigy3))
```

3. El tamaño de muestra óptimo basado en el criterio ANOREL para este caso multiparamétrico sería de $n^* = 108$. El siguiente código así lo comprueba.

```
> a1 <-  sigy1^2/(sum(((1/pik1)-1)*sigy1^2))
> a2 <-  sigy2^2/(sum(((1/pik2)-1)*sigy2^2))
> a3 <-  sigy3^2/(sum(((1/pik3)-1)*sigy3^2))
> aqk <- a1+a2+a3

> n.st <- ((sum(sqrt(aqk)))^2)/((1+0.03)*3+(sum(aqk)))
> n.st <- as.integer(n.st)
> n.st
[1] 108
```

4. El vector de probabilidades de inclusión óptimas para las tres características de interés está dado por el siguiente código. Nótese que la suma de estas en la población equivale al tamaño de muestra.

```
> pikopt <- n.st*sqrt(aqk)/sum(sqrt(aqk))
> sum(pikopt) == n.st
[1] TRUE
```

5. Como las entradas del vector de probabilidades de inclusión resultante son desiguales, entonces se debe seleccionar la muestra con algún diseño de muestreo de orden (probabilidades de inclusión desiguales y tamaño de muestra fijo). La función `UPopips` del paquete `sampling` selecciona una muestra con las anteriores características. Una vez que la muestra se selecciona, se utiliza la función `getdata` para extraer los datos observados.

```
> sam <- UPopips(pikopt,"exponential")
> getdata(MU281,sam)
  LABEL P85 P75 RMT85 CS82 SS82 S82 ME84 REV84 REG CL
     74  17  18   113    8   20  49  784  1733   3 13
    184  12  11    82    6   25  41  646   935   6 33
    ...
```

Cuando la muestra es seleccionada, el estadístico se enfrenta al problema de la estimación multi-paramétrica sobre las características de interés. Es posible escribir un código computacional para lograr la estimación de los parámetros de interés (forma tradicional) o escribir un código computacional una sola vez, basado en el enfoque matricial. Para el ejemplo de la población MU281, para la cual se obtuvieron probabilidades de inclusión óptimas, $\pi_{(opt)k}$, el estimador de Horvitz-Thompson para el vector de totales de las características de interés, para el vector de totales de las características de información auxiliar y para el tamaño poblacional se calcula mediante el siguiente código.

```
> Ys <- cbind(Y1,Y2,Y3)[sam,]
> Xs <- cbind(1,X1,X2)[sam,]
> PI <- diag(pikopt[sam])
> ones <- rep(1,n.st)

> TyHT <- t(Ys)%*%solve(PI)%*%ones
> TxHT <- t(Xs)%*%solve(PI)%*%ones
> NHT <- t(ones)%*%solve(PI)%*%ones
```

El resultado de la ejecución del anterior código es un vector de totales estimados. En particular, la estimación de los totales de las características de interés está dado por

```
> TyHT
          [,1]
Y1   6603.514
Y2  49078.942
Y3 719565.860
```

Si uno o varios dominios de interés están involucrados en la etapa de estimación, el enfoque matricial da un método simple, pero exhaustivo y efectivo, de estimación. El dominio de interés para este caso en particular corresponde a la variable `REG` la cual contiene 8 categorías geográficas. Entonces, es posible obtener estimaciones de los parámetros de interés discriminadas por estas regiones. Con el uso de la función `disjunctive` del paquete `sampling`, es posible crear la matriz de indicadores para los dominios dada en (13.3.7) y obtener las estimaciones correspondientes a (13.3.11) y (13.3.13).

```
> Z <- disjunctive(Z)[sam,]
> NdHT <- t(Z)%*%solve(PI)%*%ones
> Ty1d <- t(Ys[,1]*Z)%*%solve(PI)%*%ones
> Ty2d <- t(Ys[,2]*Z)%*%solve(PI)%*%ones
> Ty3d <- t(Ys[,3]*Z)%*%solve(PI)%*%ones
```

También es posible reunir los resultados de las estimaciones por medio de una sencilla tabla de datos dada por:

```
> TydHT <- data.frame(NdHT,Ty1d,Ty2d,Ty3d)
> TydHT
   NdHT     Ty1d      Ty2d       Ty3d
1  24.17  1045.85   8711.74  109849.38
2  51.50   895.17   6707.68   90645.23
3  29.86   594.90   4449.43   60348.27
4  49.13  1072.34   7235.20  104268.22
5  64.01  1394.64   9930.63  142899.69
6  40.50   730.81   5521.88   78599.09
7   6.44   207.89   1560.50   22637.41
8  51.79   661.88   4961.85  110318.56
```

Si el estadístico sospecha que es posible utilizar un enfoque de inferencia asistido por modelos de superpoblación, entonces se deben establecer las relaciones entre las características de información auxiliar y las características de interés por medio de un modelo. En este ejemplo particular, existen tres modelos, ξ_q $(q = 1, 2, 3)$, involucrados en un modelo general ξ. La relación es como lo dicta la siguiente expresión:

$$Y_q = \beta_{q0} + \beta_{q1}X_1 + \beta_{q2}X_2 + \boldsymbol{\varepsilon}_i \qquad q = 1, 2, 3. \tag{13.5.5}$$

Nótese que $E_{\xi_i}(\boldsymbol{\varepsilon}_i) = 0$ y que la estructura de varianza de los modelos anteriores es inducida por el paso número dos del diseño de Holmberg que en particular está dado por

$$\begin{aligned}
\Sigma_1 &= \sigma_1^2 diag(x_{11}, x_{12}, \ldots, x_{1N})^{1.4}\\
\Sigma_2 &= \sigma_2^2 diag(x_{11}, x_{12}, \ldots, x_{1N})^{2}\\
\Sigma_3 &= \sigma_3^2 \mathbf{I}_{N\times N}
\end{aligned}$$

Entonces, el modelo general toma la siguiente forma

$$\begin{pmatrix} Y_{11} & Y_{21} & Y_{31}\\ Y_{12} & Y_{22} & Y_{32}\\ \vdots & \vdots & \vdots\\ Y_{1N} & Y_{2N} & Y_{3N} \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & X_{21}\\ 1 & X_{12} & X_{22}\\ \vdots & \vdots & \vdots\\ 1 & X_{1N} & X_{2N} \end{pmatrix} \begin{pmatrix} \beta_{10} & \beta_{20} & \beta_{30}\\ \beta_{11} & \beta_{21} & \beta_{31}\\ \beta_{12} & \beta_{22} & \beta_{32} \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} & \varepsilon_{21} & \varepsilon_{31}\\ \varepsilon_{12} & \varepsilon_{22} & \varepsilon_{32}\\ \vdots & \vdots & \vdots\\ \varepsilon_{1N} & \varepsilon_{2N} & \varepsilon_{3N} \end{pmatrix} \tag{13.5.6}$$

De esta manera, la estimación de la matriz de coeficientes de regresión en la población finita, que involucra la estructura de varianza de cada modelo, dada en (13.4.7) se calcula mediante el siguiente código:

```
> A1 <- diag(pikopt[sam]*Xs[,2]^(1.4))
> B1 <- (solve(t(Xs)%*%A1%*%Xs))%*%(t(Xs)%*%A1%*%Ys[,1])
> A2 <- diag(pikopt[sam]*Xs[,2]^(2))
> B2 <- (solve(t(Xs)%*%A2%*%Xs))%*%(t(Xs)%*%A2%*%Ys[,2])
> A3 <- diag(pikopt[sam])
> B3 <- (solve(t(Xs)%*%A3%*%Xs))%*%(t(Xs)%*%A3%*%Ys[,3])

> B <- matrix(c(B1,B2,B3),ncol=3,nrow=3)
> B
            [,1]        [,2]       [,3]
[1,] -1.20582067 -25.5012341  868.01938
[2,]  1.05356147   8.3134903  104.90848
[3,]  0.01756820   0.4836855  -15.78483
```

El siguiente paso es implementar el estimador múltiple de regresión general para los totales de interés dado (13.4.10). El código computacional requiere sólo de una linea para la realización del cálculo como se muestra a continuación.

```
> TyMgreg <- TyHT+t(B)%*%(Tx - TxHT )
> TyMgreg
         [,1]
Y1   7079.411
Y2  53028.236
Y3 750689.737
```

Este estimador puede tomar distintas formas. Entre otras, puede ser reescrito de forma simplificada como en (13.4.13). Sin embargo, es necesario calcular antes la matriz de calibración dada por la expresión (13.4.14). el siguiente código muestra la implementación de la teoría

```
> w1 <- solve(PI)%*%ones + (A1%*%Xs)%*%(solve(t(Xs)%*%A1%*%Xs))%*%(Tx - TxHT)
> w2 <- solve(PI)%*%ones + (A2%*%Xs)%*%(solve(t(Xs)%*%A2%*%Xs))%*%(Tx - TxHT)
> w3 <- solve(PI)%*%ones + (A3%*%Xs)%*%(solve(t(Xs)%*%A3%*%Xs))%*%(Tx - TxHT)
> W <- cbind(w1,w2,w3)
> TyMgreg <- t(W*Ys)%*%ones
> TyMgreg
         [,1]
Y1   7079.411
Y2  53028.236
Y3 750689.737
```

El principio de calibración mostrado en (13.4.16) puede ser verificado fácilmente para cada columna de la matriz de calibración. Particularmente para la segunda columna el resultado se mantiene.

```
> t(w2)%*%Xs
          X1    X2
[1,] 281 6818 13257
```

De esta manera, se ha mostrado cómo planear y desarrollar una encuesta multi-propósito; en primera instancia, usando el diseño de muestreo de Holmberg en la etapa de diseño y el enfoque matricial en la etapa de estimación.

13.6 Marco y Lucy

Al momento de planear una encuesta, la forma tradicional se enfoca en una sola variable de interés, la cual es insuficiente para aquel estadístico que debe responder por la estimación de varios parámetros de interés. En este capítulo, e indirectamente a lo largo de todo el libro, se planteó un enfoque útil para la estimación simultanea de varios parámetros de interés. Además de las ventajas computacionales, este enfoque matricial sirve como vehículo para la introducción de tópicos de muestreo avanzados como el sistema general de ponderación propuesto por Lavallée & Caron (2008).

Por supuesto, este capítulo cierra con Marco y Lucy que, indirectamente a lo largo de todo el libro, han demostrado que el enfoque matricial de estimación simultanea debe ser usado por el estadístico teórico y práctico. Suponga que el marco de muestreo tiene la cualidad de proporcionar, además de la identificación y ubicación de cada empresa, una característica de información auxiliar como el Ingreso de cada empresa. En este orden de ideas, el lector, que ha seguido una lectura directa del libro hasta esta etapa, sabrá que el tamaño de la población es $N = 2396$ y que se han obtenido excelentes resultados con diseños de muestreo proporcionales al tamaño del Ingreso de la empresa para las características de interés Empleados e Ingreso. Además, estos buenos resultados se han obtenido con un tamaño de muestra $n = 400$.

Por otra parte, suponga que la relación entre la característica de información auxiliar Ingreso es lineal para la característica de interés Empleados pero cuadrática para la característica de interés Impuestos. Estas características se deben definir en el entorno computacional, determinando así las cantidades σ_{qk} de la expresión (13.5.1), de la siguiente manera.

```
> data(Lucy)
> attach(Lucy)
> N <- dim(Lucy)[1]
> n <- c(400,400)
> sigy1 <- sqrt(Income^(1))
> sigy2 <- sqrt(Income^(2))
> sigma <- cbind(sigy1,sigy2)
```

Recurriendo a la función `PikHol`, del paquete `TeachingSampling`, la cual contiene tres parámetros computacionales: el primero, `n`, es un vector de tamaños de muestra según la optimalidad de cada diseño para cada variable de interés involucrada en la encuesta, `sigma` una matriz, de N filas y tantas columnas como características de interés, en la cual se guardan cada una de las cantidades σ_{qk} que determinan las relaciones de las características de interés con la información auxiliar y por último, `e`, que corresponde al error máximo permitido bajo el criterio ANOREL. El resultado de la función es un vector de probabilidades de inclusión óptimas para todos los individuos de la población finita, cuya suma da como resultado el tamaño de muestra óptimo bajo este criterio ANOREL.

```
> pis <- PikHol(n,sigma,e=0.03)
> sum(pis)
[1] 398
```

De aquí en adelante, todo se torna familiar puesto que se recurre a la función `S.piPS`, del paquete `TeachingSampling`, para seleccionar una muestra aleatoria de empresas. El resultado de esta función es, por un lado, un vector conteniendo la muestra realizada y, por otro, un vector de probabilidades de inclusión de las empresas seleccionadas. Después de la recolección de los datos, se utiliza la función `E.piPS` para obtener las estimaciones que resultan óptimas bajo el criterio ANOREL.

```
> res <- S.piPS(398,pis)
> sam <- res[,1]
> Pik.s <- res[,2]
> muestra <- Lucy[sam,]
> attach(muestra)

> estima <- data.frame(Income, Employees, Taxes)
> E.piPS(estima,Pik.s)
                 Income    Employees        Taxes
Estimation 1.034870e+06 1.487223e+05 2.843080e+04
Variance   4.553146e+07 1.109766e+07 6.954306e+05
CVE        6.520335e-01 2.239957e+00 2.933174e+00
```

Los resultados del muestreo de Holmberg son óptimos para todas las variables de interés de la población de empresas del sector industrial. Son óptimos puesto que las probabilidades de inclusión que utilizó el estimador de Horvitz-Thompson fueron construidas con base en diseño óptimos para cada una de las características de interés. Además de tener la características de optimalidad, son muy precisos y sólo basta con observar el cuadro de salida para concluir que sí es posible planear una encuesta multi-propósito con diseños de muestreo que contemplen gran cantidad de características de información auxiliar y que los resultados son muy eficientes.

Tabla 13.1: *Muestreo de Holmberg: estimación de los totales de las características de interés.*

Variable	Total poblacional	Total estimado	cve %	Desv. %
Ingreso	1035217	1034870	0.65	-0.00
Empleados	151950	148722	2.23	-0.02
Impuestos	28654	28430	2.93	-0.00

13.7 Ejercicios

13.1 Demuestre la siguiente igualdad

$$Cov(\hat{t}_{y,\pi}, \hat{t}_{x,\pi}) = \sum_U \sum_U \Delta_{kl} \frac{y_k}{\pi_k} \frac{x_l}{\pi_l}$$

13.2 Demuestre que, para un diseño de muestreo aleatorio simple, se tiene la siguiente relación

$$Cov(\hat{t}_{y,\pi}, \hat{t}_{x,\pi}) = \frac{N^2}{n} \left(1 - \frac{n}{N}\right) Cov_S(y, x)$$

En donde

$$Cov_S(y,x) = \frac{1}{n-1}\sum_S (y_k - \bar{y}_S)(x_k - \bar{x}_S)$$

13.3 Obtenga una expresión para las probabilidades de inclusión de Holmberg cuando todas las medidas de tamaño son constantes en un estudio multipropósito.

13.4 Demuestre la expresión (13.3.13)

Capítulo 14

Muestreo indirecto

> En investigaciones sociales o económicas es útil contar con un marco de muestreo [en el proceso] de selección de muestras... Desafortunadamente, [a veces] sucede que no se cuenta con una lista que contenga las unidades objetivo, sino que se cuenta con otra lista de unidades que están vinculadas de cierta manera a la lista de las unidades objetivo.
>
> Jean-Claude Deville (2006)

En la práctica los marcos de muestreo de elementos no están siempre disponibles. Sin embargo, en algunas ocasiones, es posible tener acceso a diferentes marcos de lista de elementos que, si bien no hacen parte de la población objetivo, sí están indirectamente relacionados con ésta. Al proceso de selección de muestras bajo las anteriores condiciones se le llama Muestreo Indirecto que está caracterizado porque la producción de estimaciones de simples totales o medias se puede volver una pesadilla para el estadístico. Para resolver este problema se apela al método de ponderación generalizada, caracterizado por su simplicidad y cuyos estimadores comparten la buena propiedad del insesgamiento, incluso bajo muestreo indirecto.

Para producir estimaciones, en investigaciones de tipo social, económico, etc., generadas mediante una estrategia de muestreo $(p(s), \hat{T}(S))$ es imprescindible el acceso, al menos de manera implícita, a un marco de muestreo de elementos de la población objetivo, denotada como U_B. Desafortunadamente, el acceso a tal marco de muestreo es, en la mayoría de ocasiones, difícil de conseguir. Sin embargo, a veces, es posible considerar la disponibilidad de un marco muestral de elementos[1] de alguna otra población U_A vinculados con los elementos de la población objetivo.

Se pretende seleccionar una muestra probabilística s_A de la población U_A para obtener estimaciones para la población U_B usando la correspondencia entre las dos poblaciones. Por ejemplo, asuma que se desean obtener estimaciones de una población de niños con la restricción de que sólo se tiene acceso a una lista de padres conteniendo la respectiva identificación y ubicación de cada uno de ellos. La población objetivo son los niños, pero es necesario seleccionar una muestra de padres para poder entrevistar a los niños.

[1] Nótese que este es un caso particular del muestreo de conglomerados si el marco de muestreo de la población U_A fuera de conglomerados

14.1 Notación

La población U_A contiene N_A unidades. Cada unidad perteneciente a la población de U_A será rotulada con la letra j. Cada unidad perteneciente a la población objetivo U_B de tamaño N_B será rotulada con la letra i. La correspondencia entre las dos poblaciones U_A y U_B pueden ser representadas por una matriz de vínculos, denotada por $\Theta_{AB} = [\theta_{ij}^{AB}]$ de tamaño $N_A \times N_B$. Los posibles valores de la matriz están dados de la siguiente manera

$$\theta_{ij}^{AB} \begin{cases} > 0, & \text{si } j \text{ está relacionado con } i; \\ = 0, & \text{en otro caso.} \end{cases} \tag{14.1.1}$$

En el ejemplo de los padres, si la matriz de vínculos está dada por

$$\Theta_{AB} = \begin{pmatrix} \theta_{11}^{AB} & \theta_{12}^{AB} & 0 \\ 0 & \theta_{22}^{AB} & 0 \\ 0 & 0 & \theta_{33}^{AB} \\ 0 & 0 & \theta_{43}^{AB} \end{pmatrix} \tag{14.1.2}$$

entonces los vínculos existentes entre las dos poblaciones serían los siguientes

- La primera pareja, dada por los elementos 1 y 2 de la población U_A tiene un hijo notado como el segundo elemento de la población U_B.
- Sin embargo, el elemento 1 de la población U_A tiene otro hijo por fuera del matrimonio y es notado como el primer elemento de la población B.
- La segunda pareja, dada por los elementos 3 y 4 de la población U_A, tiene sólo un hijo notado como el tercer y último elemento de la población U_B

Usualmente, cuando existe un vínculo entre el elemento j-ésimo de la población U_A y el i-ésimo elemento de la población U_B, θ_{ij}^{AB} toma el valor uno. Aunque el vínculo puede ser distinto de uno como es discutido en (Lavallée 2007).

Usando muestreo indirecto, una muestra s_A de tamaño n_A es seleccionada (realizada) mediante el uso de un diseño muestral $p_A(s_A)$. Sean $\pi_j^A > 0 \ \forall j \in U_A$la probabilidad de inclusión del j-ésimo elemento. Para cada elemento en la muestra s_A se identifican las unidades en U_B cuya correspondencia con los elementos de la población U_A es no nula, es decir tales que $\theta_{ij}^{AB} > 0$. Sea s_B el conjunto de n_B unidades de la población objetivo que se lograron identificar con ayuda de los elemento pertenecientes a la población U_A. Por tanto[2]

$$s_B = \{i \in B \mid \exists j \in s_A \text{ y } \theta_{ij}^{AB} > 0\} \tag{14.1.3}$$

[2]Nótese que el conjunto s_B, aunque constituye una muestra al azar, no constituye una muestra aleatoria o probabilística puesto que su probabilidad de selección es desconocida. Se debe tener en cuenta que a pesar de que existe un diseño de muestreo $p_A(\cdot)$ que gobierna la selección aleatoria de la muestra s_A, éste no es el mismo que gobierna la selección del conjunto s_B, puesto que para dos muestras distintas, digamos s_{A1} y s_{A2}, el conjunto de unidades finales en la población U_B puede resultar el mismo. Sin embargo, de aquí en adelante llamaremos, abusando del lenguaje, al conjunto s_B como la muestra de la población U_B.

Para cada elemento identificado en la población objetivo se realiza el proceso de medición de la característica de interés y. Sin embargo, el número de elementos de la población objetivo identificadas por el proceso de muestreo indirecto es generalmente aleatorio porque no sólo depende de la muestra seleccionada s_A sino que también de la matriz de vínculo Θ_{AB}. Así que se torna muy complicado establecer un presupuesto para la etapa de recolección de la información; afortunadamente, en algunas poblaciones (como la de los padres e hijos) es posible predecir el número de vínculos entre las poblaciones (por ejemplo, un padre tiene uno, dos o incluso tres hijos).

Un requisito importante, a la hora de aplicar el muestreo indirecto, es que para todas las unidades seleccionadas en al muestra s_A se puede obtener la correspondencia a la población objetivo y viceversa. Este es un supuesto muy fuerte, aunque necesario. Por ejemplo, es fácil que un padre pueda identificar a todos sus hijos, por otro lado no es tan sencillo que un niño muy joven pueda identificar a sus padres divorciados. Sin embargo, este problema operativo se considera despreciable en términos del desarrollo teórico. De tal forma que es posible conocer los valores de la matriz Θ_{AB} para las filas $j \in s_A$ así como también para las columnas $i \in s_B$.

14.2 Estimación del total

El objetivo es estimar el total de y en la población objetivo

$$t_y = \sum_{i \in U_B} y_i \tag{14.2.1}$$

$$= \mathbf{1}_B \mathbf{y} \tag{14.2.2}$$

donde $\mathbf{1}_B$ es el vector de unos de tamaño N_B y $\mathbf{y} = (y_1, \ldots, y_{N_B})'$. Ahora, se tiene la siguiente definición que coadyuvará en la estimación del total poblacional.

Definición 14.2.1. *La matriz de vínculo estándar se define como*

$$\tilde{\Theta}_{AB} = \Theta_{AB}[diag(\mathbf{1}_A \Theta_{AB})]^{-1} \tag{14.2.3}$$

Con base en lo anterior, nótese que

$$\mathbf{1}'_A \Theta_{AB} = (\theta^{AB}_{+1}, \theta^{AB}_{+2}, \ldots, \theta^{AB}_{+N_B}) \tag{14.2.4}$$

donde $\theta^{AB}_{+i} = \sum_{j \in U_A} \theta^{AB}_{ji}$ *debe ser no nula*[3] *para todo* $i \in U_B$. *Con esto (14.2.3) está bien definida y por lo tanto* $\tilde{\Theta}^{AB}_{ji} = \dfrac{\theta^{AB}_{ji}}{\theta^{AB}_{+i}}$.

En la población de ejemplo, significaría que todo hijo debe estar vinculado al menos a un padre, lo cual es lógico en este contexto específico. Sin embargo esta lógica no siempre se cumple y en algunas ocasiones la definición propia de la población U_A es compleja.

[3]Esta restricción indica que todos los miembros de la población U_A deben tener al menos un vínculo con algún individuo de la población objetivo. Aún más, con esta restricción, si existe algún miembro de la población U_A que no tenga vínculo con algún miembro de la población U_B no debe ser considerado.

Resultado 14.2.1. *Si $\tilde{\Theta}_{AB}$ es una matriz de vínculo estándar, entonces*

$$\tilde{\Theta}'_{AB}\mathbf{1}_A = \mathbf{1}_B \tag{14.2.5}$$

Prueba. Desarrollando algebraicamente, se tiene la demostración directamente al aplicar la anterior definición, de la siguiente manera:

$$\begin{aligned}
\tilde{\Theta}'_{AB}\mathbf{1}_A &= \left([\text{diag}(\mathbf{1}_A\Theta_{AB})]^{-1}\right)'\Theta'_{AB}\mathbf{1}_A \\
&= [\text{diag}(\mathbf{1}_A\Theta_{AB})]^{-1}(\mathbf{1}'_A\Theta_{AB})' \\
&= \begin{pmatrix} \frac{1}{\theta^{AB}_{+1}} & 0 & \cdots & 0 \\ 0 & \frac{1}{\theta^{AB}_{+2}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\theta^{AB}_{+N_B}} \end{pmatrix} \begin{pmatrix} \theta^{AB}_{+1} \\ \theta^{AB}_{+2} \\ \vdots \\ \theta^{AB}_{+N_B} \end{pmatrix} = \mathbf{1}_B
\end{aligned}$$

■

Resultado 14.2.2. *El total poblacional de la característica de interés puede ser reescrito de la siguiente manera*

$$t_y = \sum_{j\in U_A}\sum_{i\in U_B}\frac{\theta^{AB}_{ji}}{\theta^{AB}_{+i}}y_i \tag{14.2.6}$$

Prueba. Directamente de la definición de la matriz de vínculo estándar, se tiene que

$$\begin{aligned}
t_y &= \mathbf{1}'_B\mathbf{y} \\
&= \mathbf{1}'_A\tilde{\Theta}_{AB}\mathbf{y} = \sum_{j\in U_A}\sum_{i\in U_B}\frac{\theta^{AB}_{ji}}{\theta^{AB}_{+i}}y_i
\end{aligned}$$

■

A continuación se define el vector columna $\mathbf{z} = \tilde{\Theta}_{AB}\mathbf{y}$ de tamaño N_A cuyo j-ésimo elemento es $z_j = \sum_{i\in U_B}\tilde{\theta}^{AB}_{ji}y_i$ establecido para la población U_A y medido en la muestra s_A. Para estimar t_y, se debe recurrir a la utilización de los valores de y_i medidos en la muestra s_B, de tal forma que es posible construir el siguiente estimador

$$\begin{aligned}
\hat{t}_y &= \sum_{i\in U_B} w_i y_i \qquad (14.2.7)\\
&= \mathbf{w}'\mathbf{y} \qquad (14.2.8)
\end{aligned}$$

donde $\mathbf{w} = (w_1, \ldots, w_{N_B})$, w_i es la ponderación estimada del i-ésimo elemento de s_B. Por supuesto, $w_i = 0$ si $i \notin s_B$. Para que $\hat{t}_y$ sea insesgado es usual definir $w_i = (\pi^B_i)^{-1}$. Sin embargo, esta escogencia, aunque posible en la teoría, es muy difícil de hallar en muestreo indirecto puesto que se debe tener conocimiento de todos los posibles vínculos generados por todas las posibles muestras a_A.

14.3 Método de ponderación generalizada

La muestra s_A fue seleccionada de acuerdo a un diseño de muestreo $p_A(s_A)$. Este diseño de muestreo induce un vector de probabilidades de inclusión para todos los elementos de U_A. Sea $\mathbf{\Pi}_A = diag(\pi_1^A, \ldots, \pi_{N_A}^A)$ una matriz diagonal de tamaño $N_A \times N_A$ conteniendo las probabilidades de inclusión para $j \in U_A$ en su diagonal. De igual manera, se define la matriz de inclusión de los elementos en la muestra dada por $\mathbf{I_A} = diag(I_1^A, \ldots, I_{N_A}^A)$ con

$$I_j^A(S_A) = \begin{cases} 1 & \text{si } i \in S_A \\ 0 & \text{si } i \notin S_A. \end{cases} \tag{14.3.1}$$

Partiendo de que el total poblacional toma la siguiente forma

$$\begin{aligned} t_y &= \mathbf{1}'_A \tilde{\Theta}_{AB} \mathbf{y} \\ &= \mathbf{1}'_A \mathbf{z} \end{aligned}$$

entonces, es posible construir una expresión que respete los principios del estimador de Horvitz-Thompson en términos del vector $\mathbf{Z}$. Por lo tanto

$$\hat{t}_y = \hat{t}_{z,\pi} = \mathbf{1}'_A \mathbf{I_A} \mathbf{\Pi}_A^{-1} \mathbf{z} \tag{14.3.2}$$

$$= \mathbf{1}'_A \mathbf{I_A} \mathbf{\Pi}_A^{-1} \tilde{\Theta}_{AB} \mathbf{y} \tag{14.3.3}$$

Por ello, se define el vector de ponderaciones para la población objetivo U_B como

$$\mathbf{w} = \mathbf{1}'_A \mathbf{I_A} \mathbf{\Pi}_A^{-1} \tilde{\Theta}_{AB} \tag{14.3.4}$$

donde cada elemento de $\mathbf{w}$, el cual es un vector de tamaño N_B, está definido por la siguiente expresión

$$w_i = \begin{cases} \sum_{j \in U_A} I_j \dfrac{\tilde{\Theta}_{ji}^{AB}}{\pi_j^A}, & \text{para todo } i \in s_B \\ 0, & \text{para todo } i \notin s_B \end{cases} \tag{14.3.5}$$

De esta forma, se dice que los pesos w_i han sido obtenidos mediante el método de ponderación generalizada tal como se describe en (Lavallée 2007). En este orden de ideas, y retomando nuestro ejemplo de la población de padres e hijos, si la muestra realizada de padres estuviera dada por

s_A=**Padre 2, Padre 3.**

entonces el conjunto de niños identificados por los padres seleccionados estaría dado por

s_B=**Niño 2, Niño 3.**

y las ponderaciones resultantes serían

- $w_1 = 0$ pues el **Niño 1** no fue identificado por ningún padre
- Para el **Niño 2** se tiene que

$$\begin{aligned} w_2 &= \sum_{j \in U_A} I_j \frac{\tilde{\Theta}^{AB}_{j2}}{\pi^A_1} \\ &= \sum_{j \in s_A} \frac{\tilde{\Theta}^{AB}_{j2}}{\pi^A_1} \\ &= \frac{\tilde{\Theta}^{AB}_{22}}{\pi^A_2} + \frac{\tilde{\Theta}^{AB}_{23}}{\pi^A_3} \\ &= \frac{\Theta^{AB}_{22}}{\Theta^{AB}_{+2}} \frac{1}{\pi^A_2} + \frac{\Theta^{AB}_{23}}{\Theta^{AB}_{+2}} \frac{1}{\pi^A_3} \end{aligned}$$

- Para el **Niño 3** se tiene que

$$\begin{aligned} w_3 &= \sum_{j \in U_A} I_j \frac{\tilde{\Theta}^{AB}_{j3}}{\pi^A_1} \\ &= \sum_{j \in s_A} \frac{\tilde{\Theta}^{AB}_{j3}}{\pi^A_1} \\ &= \frac{\tilde{\Theta}^{AB}_{32}}{\pi^A_2} + \frac{\tilde{\Theta}^{AB}_{33}}{\pi^A_3} \\ &= \frac{\Theta^{AB}_{32}}{\Theta^{AB}_{+3}} \frac{1}{\pi^A_2} + \frac{\Theta^{AB}_{33}}{\Theta^{AB}_{+3}} \frac{1}{\pi^A_3} \end{aligned}$$

14.3.1 Propiedades

Se tienen las siguientes propiedades generadas de los pesos del método de ponderación generalizada

Resultado 14.3.1. *El estimador t_y es insesgado*

Prueba. Basta con demostrar que $E(\mathbf{w}) = \mathbf{1}_B$. Esto se tiene por construcción, dado que el estimador de Horvitz-Thompson es insesgado puesto que la esperanza de las variables indicadoras I_j es igual a la probabilidad de inclusión π^A_j. De esta manera, se tiene que

$$\begin{aligned} E(\hat{t}_y) &= E(\mathbf{w})y \\ &= E(\mathbf{1}'_A \mathbf{I_A} \mathbf{\Pi}^{-1}_A \tilde{\Theta}_{AB})y \\ &= \mathbf{1}'_A E(\mathbf{I_A}) \mathbf{\Pi}^{-1}_A \tilde{\Theta}_{AB} y \\ &= \mathbf{1}'_A \mathbf{\Pi}_A \mathbf{\Pi}^{-1}_A \tilde{\Theta}_{AB} y \\ &= \mathbf{1}'_A \tilde{\Theta}_{AB} y \\ &= \mathbf{1}'_B y = t_y \end{aligned}$$



Resultado 14.3.2. *El vector* $\mathbf{w}$ *provee estimaciones insesgadas si y sólo si la matriz* $\tilde{\Theta}_{AB}$ *es una matriz de vínculo estándar*

Prueba. Se tiene que $E(\mathbf{w}) = \tilde{\Theta}'_{AB}\mathbf{1}_A$; sin embargo asumiendo, por el resultado anterior, que el vector de pesos induce estimaciones insesgadas tenemos que $E(\mathbf{w}) = \mathbf{1}_B$. Por tanto $\tilde{\Theta}'_{AB}\mathbf{1}_A = \mathbf{1}_B$ y, con base en este razonamiento, se tiene la demostración del resultado. ■

Resultado 14.3.3. *La varianza de* $\hat{t}_y$ *está dada por*

$$
\begin{aligned}
Var(\hat{t}_y) &= \mathbf{z}'\mathbf{\Delta}_A\mathbf{z} \qquad (14.3.6)\\
&= \mathbf{y}'\mathbf{\Delta}_B\mathbf{y} \qquad (14.3.7)
\end{aligned}
$$

donde $\mathbf{\Delta}_B = \tilde{\Theta}'_{AB}\mathbf{\Delta}_A\tilde{\Theta}_{AB}$ *y* $\mathbf{\Delta}_A$ *es la matriz de varianzas y covarianzas de tamaño* $N_A \times N_A$ *de las variables indicadoras de los elementos de la población* U_A *doblemente ponderada por probabilidades de inclusion cuyo elemento* jj' *está dado por*

$$
[\mathbf{\Delta}_A]_{jj'} = \frac{\Delta^A_{jj'}}{\pi^A_j\pi^A_{j'}} = \frac{\pi^A_{jj'} - \pi^A_j\pi^A_{j'}}{\pi^A_j\pi^A_{j'}}
$$

Prueba. Siguiendo los principios del estimador de Horvitz-Thompson se tiene la demostración de manera inmediata, puesto que

$$
\mathbf{z}'\mathbf{\Delta}_A\mathbf{z} = (z_1,\ldots,z_{N_A})\begin{pmatrix} \frac{\Delta^A_{11}}{\pi^A_1\pi^A_1} & \cdots & \frac{\Delta^A_{1N_A}}{\pi^A_1\pi^A_{N_A}} \\ \vdots & \ddots & \vdots \\ \frac{\Delta^A_{N_A1}}{\pi^A_{N_A}\pi^A_1} & \cdots & \frac{\Delta^A_{N_AN_A}}{\pi^A_{N_A}\pi^A_{N_A}} \end{pmatrix}\begin{pmatrix} z_1 \\ \vdots \\ z_{N_A} \end{pmatrix} \qquad (14.3.8)
$$

$$
= \left(\sum_{j\in U_A} z_j\frac{\Delta^A_{j1}}{\pi^A_1\pi^A_j},\ldots,\sum_{j\in U_A} z_j\frac{\Delta^A_{jN_A}}{\pi^A_{N_A}\pi^A_j}\right)\begin{pmatrix} z_1 \\ \vdots \\ z_{N_A} \end{pmatrix} \qquad (14.3.9)
$$

$$
= \sum_{j\in U_A}\sum_{j'\in U_A}\Delta_{jj'}\frac{z_j}{\pi^A_j}\frac{z_{j'}}{\pi^A_{j'}} = Var(\hat{t}_{z,\pi}) \qquad (14.3.10)
$$

y reemplazando convenientemente se obtiene la demostración. ■

14.3.2 Algunas matrices especiales

En general, el muestreo indirecto produce estimaciones insesgadas si se utiliza el método de ponderación generalizada. Sin embargo, vale la pena presentar casos especiales de matrices de vínculo que ilustren el comportamiento del estimador de Horvitz-Thompson. En este apartado, se presentan algunas de estas matrices que corresponden a casos extremos, que aunque posiblemente no sean plausibles en la práctica, sirven para ilustrar el efecto de la matriz de vínculo sobre el estimador del total poblacional de la característica de interés.

Matriz identidad

Al asumir que la matriz de vínculo es una matriz identidad, se tiene que la población U_A y la población U_B tienen una correspondencia uno a uno. Esto implica que el tamaño de las dos poblaciones es el mismo, así $N_A = N_B = N$ y que la matriz de vínculo está dada por

$$\Theta_{AB} = \mathbf{I}_{N\times N} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}$$

Con lo que el vector de pesos es

$$\mathbf{w} = \left(\frac{I_1^A}{\pi_1^A}, \cdots, \frac{I_{N_A}^A}{\pi_{N_A}^A}\right)' \tag{14.3.11}$$

y por tanto $\mathbf{Z} = \mathbf{y}$. Luego, el estimador $\hat{t}_y$ tomará la forma del estimador de Narain-Horvitz-Thompson así

$$\hat{t}_y = \hat{t}_{y,N\pi} = \mathbf{1}'_A \mathbf{I_A} \mathbf{\Pi}_A^{-1} \mathbf{y} \tag{14.3.12}$$

Uno para todos

Considere el caso en que la población objetivo se encuentra particionada en N_I^B conglomerados, cada uno de tamaño N_i^B $i = 1, \ldots, N_I^B$. Cada conglomerados de U_I^B está asociado exactamente con un elemento j de U_A. Nótese que $N_I^B = N^A$. Por tanto, la matriz de vínculo está dada por

$$\Theta_{AB} = \begin{pmatrix} \mathbf{1}'_{B1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}'_{B2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{1}'_{BN_I^B} \end{pmatrix} \tag{14.3.13}$$

donde $\mathbf{1}_{Bi}$ el vector de unos de tamaño N_i^B $i = 1, \ldots, N_I$. La matriz de vínculo también puede ser escrita como $\Theta_{AB} = \text{diag}(\mathbf{1}'_{B1}, \ldots, \mathbf{1}'_{BN_I^B}$ por tanto la matriz de vínculo estandarizada toma la siguiente forma

$$\begin{aligned} \tilde{\Theta}_{AB} &= \Theta_{AB}[\text{diag}(\mathbf{1}_A \Theta_{AB})]^{-1} & (14.3.14) \\ &= \Theta_{AB}[\text{diag}(\mathbf{1}_A \text{diag}(\mathbf{1}'_{B1}, \ldots, \mathbf{1}'_{BN_I^B}))]^{-1} & (14.3.15) \\ &= \Theta_{AB}[\text{diag}(\mathbf{1}'_{B1}, \ldots, \mathbf{1}'_{BN_I^B})]^{-1} & (14.3.16) \\ &= \Theta_{AB}[\mathbf{I}_{\sum_{i=1}^{N_i} N_i^B \times \sum_{i=1}^{N_i} N_i^B}]^{-1} & (14.3.17) \\ &= \Theta_{AB} & (14.3.18) \end{aligned}$$

Se tiene entonces que el vector de pesos $\mathbf{w}$ está definido como

$$\mathbf{w} = \left(\frac{I_1^A}{\pi_1^A}\mathbf{1}_{B1}, \ldots, \frac{I_{N_I^A}^A}{\pi_1^A}\mathbf{1}_{BN_I^B}\right)' \tag{14.3.19}$$

y el estimador se puede escribir como

$$\hat{t}_y = \sum_{i=1}^{N_I^B} \frac{I_i^A}{\pi_i^A} t_{y,U_i} \tag{14.3.20}$$

donde $t_{y,U_i} = \sum_{k \in U_i^B} y_k$ es el total de i-ésimo conglomerado de la población U_I^B

Todos para uno

En este caso se considera que la población U^A está particionada en N_I^A conglomerados, cada uno de tamaño N_j^A $j = 1, \ldots, N_I^A$. Cada conglomerado de U_I^A está asociado exactamente con un elemento i de U_B. Nótese que $N_I^A = N^B$. Por tanto, la matriz de vínculo está dada por

$$\Theta_{AB} = \begin{pmatrix} \mathbf{1}_{A1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{A2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{1}'_{AN_I^A} \end{pmatrix} \tag{14.3.21}$$

donde $\mathbf{1}_{Aj}$ el vector de unos de tamaño N_i^A $j = 1, \ldots, N_I^A$. En este caso particular, la matriz de vínculo estandarizada está dada por la siguiente expresión

$$\tilde{\Theta}_{AB} = \Theta_{AB}[\text{diag}(\mathbf{1}_A \Theta_{AB})]^{-1} \tag{14.3.22}$$

$$= \text{diag}(\frac{1}{N_1^A}\mathbf{1}_{A1}, \ldots, \frac{1}{N_{N_I^A}^A}\mathbf{1}_{AN_I^A}) \tag{14.3.23}$$

Se tiene entonces que el vector de pesos $\mathbf{w}$ está definido como

$$\mathbf{w} = \left(\frac{1}{N_1^A} \sum_{j \in U_I^A} \frac{I_j^A}{\pi_j^A}, \ldots, \frac{1}{N_{N_I^A}^A} \sum_{j \in U_{N_I^A}^A} \frac{I_j^A}{\pi_j^A} \right)' \tag{14.3.24}$$

y el estimador resultante toma la siguiente forma

$$\hat{t}_y = \sum_{i=1}^{N_I^A} \frac{y_i}{N_i^A} \sum_{j \in U_i^A} \frac{I_j^A}{\pi_j^A} \tag{14.3.25}$$

14.4 Ejemplo léxico-gráfico

Suponga que en el ejemplo de la población de Padres e Hijos, cuya matriz de vínculos está dada por la expresión (14.1.2), y se plantea una investigación acerca de la estimación del total de los Hijos. Para tales efectos, se supone que la matriz de vínculos, siguiendo el consejo de Lavallée (2007) está dada por

$$\Theta_{AB} = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \tag{14.4.1}$$

En resumen, la población de los Padres U_A está compuesta por $N_A = 4$ individuos. Se desea seleccionar una muestra S_A de $n_A = 2$ individuos mediante un diseño de muestreo aleatorio simple. Bajo esta configuración existen $\binom{4}{2} = 6$ posibles muestras. Por otro lado, la población objetivo U_B está compuesta por $N_B = 3$ Hijos cuyas edades son 2 años, 3 años y 3 años, en estricto orden; es decir, el total poblacional de la característica de interés es $2 + 3 + 3 = 8$ años. Repasemos cada una de las posible muestras y veamos que, efectivamente, el estimador resultante es insesgado para el total poblacional.

Con esta configuración, es necesario hallar la matriz de vínculo estándar poblacional. De esta manera de la definición 14.2.1, se tiene que

$$\theta_{+1}^{AB} = 1, \quad \theta_{+2}^{AB} = 2, \quad \theta_{+3}^{AB} = 2$$

Por lo tanto, la matriz de vínculo estándar está dada por

$$\tilde{\Theta}_{AB} = \begin{pmatrix} 1 & 1/2 & 0 \\ 0 & 1/2 & 0 \\ 0 & 0 & 1/2 \\ 0 & 0 & 1/2 \end{pmatrix} \tag{14.4.2}$$

- **Primera muestra**: $s_A = \{Padre1, Padre2\}$. El Padre1 vincula al Hijo1 y al Hijo2, mientras que el Padre2 solamente vincula al Hijo2. De esta manera queda definida la muestra de la población objetivo como $s_B = \{Hijo1, Hijo2\}$. Las ponderaciones están dadas a continuación.

  1. $w_1 = \sum_{j\in s_A} \frac{\tilde{\theta}_{j1}^{AB}}{\pi_j^A} = (4/2)\sum_{j\in s_A} \tilde{\theta}_{j1}^{AB} = 2(1+0) = 2$
  2. $w_2 = \sum_{j\in s_A} \frac{\tilde{\theta}_{j2}^{AB}}{\pi_j^A} = (4/2)\sum_{j\in s_A} \tilde{\theta}_{j2}^{AB} = 2(1/2+1/2) = 2$
  3. $w_3 = 0$ puesto que ningún Padre lo vinculó.

  Después de recolectadas las observaciones, el vector de valores para la característica de interés es $y_s = (2, 3)$ se tiene que la estimación es $\hat{t}_y = \sum_{i\in s_B} w_i y_i = (2\times 2) + (2\times 3) = 10$

- **Segunda muestra**: $s_A = \{Padre1, Padre3\}$. El Padre1 vincula al Hijo1 y al Hijo2, mientras que el Padre3 solamente vincula al Hijo3. De esta manera queda definida la muestra de la población objetivo como $s_B = \{Hijo1, Hijo2, Hijo3\}$. Las ponderaciones están dadas a continuación.

  1. $w_1 = \sum_{j\in s_A} \frac{\tilde{\theta}_{j1}^{AB}}{\pi_j^A} = (4/2)\sum_{j\in s_A} \tilde{\theta}_{j1}^{AB} = 2(1+0) = 2$
  2. $w_2 = \sum_{j\in s_A} \frac{\tilde{\theta}_{j2}^{AB}}{\pi_j^A} = (4/2)\sum_{j\in s_A} \tilde{\theta}_{j2}^{AB} = 2(1/2+0) = 1$
  3. $w_3 = \sum_{j\in s_A} \frac{\tilde{\theta}_{j3}^{AB}}{\pi_j^A} = (4/2)\sum_{j\in s_A} \tilde{\theta}_{j3}^{AB} = 2(0+1/2) = 1$

  Después de recolectadas las observaciones, el vector de valores para la característica de interés es $y_s = (2, 3, 3)$ se tiene que la estimación es $\hat{t}_y = \sum_{i\in s_B} w_i y_i = (2\times 2) + (1\times 3) + (1\times 3) = 10$

- **Tercera muestra**: $s_A = \{Padre1, Padre4\}$. El Padre1 vincula al Hijo1 y al Hijo2, mientras que el Padre4 solamente vincula al Hijo3. De esta manera queda definida la muestra de la población objetivo como $s_B = \{Hijo1, Hijo2, Hijo3\}$. Las ponderaciones están dadas a continuación.

  1. $w_1 = \sum_{j \in s_A} \frac{\tilde{\theta}_{j1}^{AB}}{\pi_j^A} = (4/2) \sum_{j \in s_A} \tilde{\theta}_{j1}^{AB} = 2(1+0) = 2$
  2. $w_2 = \sum_{j \in s_A} \frac{\tilde{\theta}_{j2}^{AB}}{\pi_j^A} = (4/2) \sum_{j \in s_A} \tilde{\theta}_{j2}^{AB} = 2(1/2+0) = 1$
  3. $w_3 = \sum_{j \in s_A} \frac{\tilde{\theta}_{j3}^{AB}}{\pi_j^A} = (4/2) \sum_{j \in s_A} \tilde{\theta}_{j3}^{AB} = 2(0+1/2) = 1$

  Después de recolectadas las observaciones, el vector de valores para la característica de interés es $y_s = (2, 3, 3)$ se tiene que la estimación es $\hat{t}_y = \sum_{i \in s_B} w_i y_i = (2 \times 2) + (1 \times 3) + (1 \times 3) = 10$

- **Cuarta muestra**: $s_A = \{Padre2, Padre3\}$. El Padre2 vincula solamente al Hijo2, mientras que el Padre4 solamente vincula al Hijo3. De esta manera queda definida la muestra de la población objetivo como $s_B = \{Hijo2, Hijo3\}$. Las ponderaciones están dadas a continuación.

  1. $w_1 = 0$ puesto que ningún Padre lo vinculó.
  2. $w_2 = \sum_{j \in s_A} \frac{\tilde{\theta}_{j2}^{AB}}{\pi_j^A} = (4/2) \sum_{j \in s_A} \tilde{\theta}_{j2}^{AB} = 2(1/2+0) = 1$
  3. $w_3 = \sum_{j \in s_A} \frac{\tilde{\theta}_{j3}^{AB}}{\pi_j^A} = (4/2) \sum_{j \in s_A} \tilde{\theta}_{j3}^{AB} = 2(0+1/2) = 1$

  Después de recolectadas las observaciones, el vector de valores para la característica de interés es $y_s = (3, 3)$ se tiene que la estimación es $\hat{t}_y = \sum_{i \in s_B} w_i y_i = (1 \times 3) + (1 \times 3) = 6$

- **Quinta muestra**: $s_A = \{Padre2, Padre4\}$. El Padre2 vincula solamente al Hijo2, mientras que el Padre4 solamente vincula al Hijo3. De esta manera queda definida la muestra de la población objetivo como $s_B = \{Hijo2, Hijo3\}$. Las ponderaciones están dadas a continuación.

  1. $w_1 = 0$ puesto que ningún Padre lo vinculó.
  2. $w_2 = \sum_{j \in s_A} \frac{\tilde{\theta}_{j2}^{AB}}{\pi_j^A} = (4/2) \sum_{j \in s_A} \tilde{\theta}_{j2}^{AB} = 2(1/2+0) = 1$
  3. $w_3 = \sum_{j \in s_A} \frac{\tilde{\theta}_{j3}^{AB}}{\pi_j^A} = (2) \sum_{j \in s_A} \tilde{\theta}_{j3}^{AB} = 2(0+1/2) = 1$

  Después de recolectadas las observaciones, el vector de valores para la característica de interés es $y_s = (3, 3)$ se tiene que la estimación es $\hat{t}_y = \sum_{i \in s_B} w_i y_i = (1 \times 3) + (1 \times 3) = 6$

- **Sexta muestra**: $s_A = \{Padre3, Padre4\}$. El Padre3 vincula solamente al Hijo3, al igual que el Padre4. De esta manera queda definida la muestra de la población objetivo como $s_B = \{Hijo3\}$. Las ponderaciones están dadas a continuación.

  1. $w_1 = 0$ puesto que ningún Padre lo vinculó.
  2. $w_2 = 0$ puesto que ningún Padre lo vinculó.

3. $w_3 = \sum_{j \in s_A} \frac{\tilde{\theta}_{j3}^{AB}}{\pi_j^A} = (2) \sum_{j \in s_A} \tilde{\theta}_{j3}^{AB} = 2(1/2 + 1/2) = 2$

Después de recolectadas las observaciones, el vector de valores para la característica de interés es $y_s = 3$ se tiene que la estimación es $\hat{t}_y = \sum_{i \in s_B} w_i y_i = (2 \times 3) = 6$

En resumen, promediando las estimaciones con respecto al diseño de muestreo p_A, se encuentra fácilmente que el estimador es insesgado puesto que

$$(1/6) \times (10 + 10 + 10 + 6 + 6 + 6) = 8 = t_y$$

Por otro lado, nótese que el diseño de muestreo para la población objetivo, el cual es desconocido siempre que se seleccione una sola muestra, está dado a continuación

$$p_B(s_B) = \begin{cases} 2/6, & \text{si } s_B = \{Hijo1, Hijo2, Hijo3\} \\ 2/6, & \text{si } s_B = \{Hijo2, Hijo3\} \\ 1/6, & \text{si } s_B = \{Hijo1, Hijo2\} \\ 1/6, & \text{si } s_B = \{Hijo3\} \end{cases} \tag{14.4.3}$$

Por supuesto, en un ejercicio ilustrativo de este estilo, sería posible calcular las probabilidades de inclusión y utilizar el estimador de Horvitz-Thompson para estimar el total poblacional. Sin embargo, en la vida práctica esta opción se descarta rápidamente a medida que se aumenta la complejidad del diseño de muestreo y el tamaño de muestra.

14.5 Ejercicios

14.1 Suponga que se requiere la estimación del total de kilovatios al mes consumidos por los hogares de un municipio. Además, asuma que no existe un marco de muestreo de hogares, aunque sí de individuos, y que para acceder a la información requerida se diseña una muestra de individuos a los que se les pregunta por la información de su hogar.

- Argumente por qué este problema puede ser resuelto con un enfoque de muestreo indirecto.
- Con base en lo anterior, proponga un estimador para el total de kilovatios consumidos en los hogares mediante el método de ponderación generalizada.
- Si se seleccionó una muestra de n individuos, defina la probabilidad de inclusión de un hogar compuesto por $M < n$ individuos.
- Con base en lo anterior escriba las expresiones teóricas de los estimadores de Horvitz-Thonmpson y de Hajek para el total de kilovatios consumidos en los hogares.

14.2 Bajo muestreo indirecto, proponga una expresión para el estimador del total poblacional, si la muestra s_A se seleccionó de forma aleatoria simple.

14.3 Bajo muestreo indirecto, proponga una expresión para el estimador del total poblacional, si la muestra s_A se seleccionó de forma aleatoria estratificada.

14.4 Bajo muestreo indirecto, proponga una expresión para el estimador del total poblacional, si la muestra s_A se seleccionó de forma bietápica.

14.5 Bajo muestreo indirecto, proponga una expresión para el estimador del total poblacional, si se utiliza un estimador general de regresión con características de información auxiliar de la población U_A.

14.6 Bajo muestreo indirecto, proponga una expresión para el estimador del total poblacional, si se utiliza un estimador general de regresión con características de información auxiliar de la población U_B.

14.7 ¿Qué formas de ausencia de respuesta se pueden presentar en muestreo indirecto?.

14.8 Discuta por qué el muestreo de redes se puede ver como un caso particular de muestreo indirecto y proponga un estimador que involucre el método de ponderación generalizada.

14.9 Discuta por qué el muestreo adaptativo se puede ver como un caso particular de muestreo indirecto y proponga un estimador que involucre el método de ponderación generalizada.

14.10 Discuta por qué el muestreo bola de nieve se puede ver como un caso particular de muestreo indirecto y proponga un estimador que involucre el método de ponderación generalizada.

Capítulo 15

Muestreo balanceado

> El método del cubo propone un procedimiento general que permite la selección de muestras aleatorias balanceadas, con probabilidades de inclusión simples o desiguales en el sentido de que las estimaciones de Horvitz-Thompson son iguales, o casi iguales, al total poblacional de las variables de balanceo.
>
> Yves Tillé (2006)

Comúnmente, el muestreo balanceado ha sido conocido como una técnica de muestreo no probabilístico tal como el muestreo por cuotas, por conveniencia o por juzgamiento. Este tipo de muestreo sugiere la selección de muestras, para las cuales la media muestral de una característica de información auxiliar sea idéntica a la media poblacional de dicha característica de información auxiliar. Es más, si esta característica de información auxiliar está bien correlacionada con la característica de interés, entonces se dice que el muestreo balanceado es óptimo puesto que reproducirá con precisión el total o la media de la característica de interés en la población.

Tillé (2006) afirma que la idea de seleccionar muestras balanceadas nació con Neyman (1934) cuando afirmó que «el método de la selección a conveniencia consiste en a) dividir la población de distritos en estratos de segundo orden de acuerdo a los valores de x e y, b) seleccionar aleatoriamente de cada estrato un número fijo de distritos. El número de selecciones está determinado por la condición del mantenimiento del promedio ponderado de la característica de interés». Más adelante, en Yates (1946) se encuentra el siguiente extracto: «Se debe seleccionar una muestra aleatoria. Los individuos serán incluidos mediante el mismo proceso aleatorio, el primer miembro será comparado con el primer miembro de la muestra original, el segundo individuo con el segundo de la muestra original y así sucesivamente. Un nuevo miembro será sustituido si mejora el balance».

Recientemente, se ha llegado a soluciones parciales para la selección aleatoria (mediante diseños de muestreo propiamente definidos) de muestras balanceadas por medio de métodos propuestos por algunos reconocidos autores de como Ardilly (1991) y Deville (1992). Por otra parte, autores como y Valliant, Dorfman & Royall (2000) o Royal & Herson (1973) han considerado la construcción de estimadores,

enmarcados bajo métodos de inferencia basada solamente en modelos poblacionales, y su optimalidad desde el punto de vista del modelo sin tomar en cuenta el diseño muestral y concluyen que un diseño de muestreo puede ser balanceado aunque no necesariamente aleatorio o probabilístico.

Por otro lado, Deville & Tillé (2004) desarrollaron un procedimiento general y riguroso que permite la extracción de muestras probabilísticas balanceadas y la posterior estimación de las cantidades de interés, enmarcados bajo métodos de inferencia basados en el diseño de muestreo. Este procedimiento es conocido como el método del cubo y permite la selección de muestras aleatorias sobre un conjunto de características de información auxiliar (o variables de balanceo), y tiene la agradable propiedad de que el estimador de Horvitz-Thompson reproduce el total poblacional de las variables de balanceo. Más adelante, Deville & Tillé (2005) adaptaron una aproximación de la varianza para el estimador de Horvitz-Thompson en muestreo balanceado.

15.1 Notación

Dado que bajo un diseño de muestreo balanceado, el estimador de Horvitz-Thompson, para los totales de un conjunto de variables auxiliares, debe ser igual al total poblacional de las mismas, la varianzas del estimador del total poblacional de la característica de interés se debe reducir de acuerdo al aumento de correlación con las variables auxiliares.

El objetivo es estimar el total poblacional de la característica de interés $t_y = \sum_{k\in U} y_k$, entonces se supone que los valores de los vectores

$$\mathbf{x}_k = (x_{k1}, x_{k2}, \ldots, x_{kQ})'$$

tomados para q variables de balanceo, se conocen para todas las unidades de la población. Por tanto, el vector de totales de las variables de balanceo

$$\mathbf{t_x} = \sum_{k\in U} \mathbf{x}'_k$$

es también conocido, y puede ser estimado, utilizando el estimador de Horvitz-Thompson, por medio de la siguiente expresión

$$\hat{\mathbf{t}}_{\mathbf{x},\pi} = \sum_{k\in U} \frac{\mathbf{x}_k}{\pi_k} I_k.$$

El objetivo es construir un diseño de muestreo balanceado, definido como sigue.

Definición 15.1.1. *Un* ***diseño de muestreo es balanceado*** *con respecto a las variables auxiliares $x_1, ..., x_Q$, sí y sólo sí éste satisface las ecuaciones de balance dadas por*

$$\hat{\mathbf{t}}_{\mathbf{x},\pi} = \mathbf{t_x} \tag{15.1.1}$$

para toda muestra $s \in \mathcal{S}$ tal que $p(\mathbf{s}) > 0$ y para todo $q = 1, ..., Q$. En otras palabras

$$Var(\hat{\mathbf{t}}_{\mathbf{x},\pi}) = \mathbf{0}$$

Nótese que $Var(\hat{\mathbf{X}}_\pi)$ es una matriz de varianzas covarianzas. En estos términos, el diseño de muestreo balanceado, define un soporte $\mathcal{Q}$ dado por

$$\mathcal{Q} = \left\{ \mathbf{I} \in \mathcal{S} | \sum_{k \in U} \frac{\mathbf{x}_k}{\pi_k} I_k = \mathbf{t_x} \right\}$$

donde $\mathbf{I} = (I_1, \ldots, I_n)'$ es el vector de inclusión de los elementos en la muestra y $\mathcal{S}$ es el soporte simétrico sin reemplazo. Para aceptar que un diseño de muestreo puede estar condicionado, el lector deberá estar familiarizado con las definiciones dadas en los primeros capítulos de este texto. En particular, nótese que de la definición 2.1.5, el soporte simétrico sin reemplazo, que permite la definición del muestreo aleatorio simple, entre otros, es también un soporte condicionado y dado por

$$\mathcal{S}_n = \left\{ \mathbf{s} \in \mathcal{S} | \sum_{k \in U} s_k = n \right\}$$

También, el soporte simétrico con reemplazo de tamaño fijo, que permite la debida definición del diseño aleatorio simple con reemplazo, entre otros, está condicionado puesto que

$$\mathcal{R}_n = \left\{ \mathbf{s} \in \mathcal{R} | \sum_{k \in U} s_k = n \right\}$$

15.1.1 Ejemplos

A continuación se presentan algunos ejemplos que, si bien no son útiles en la práctica, sí ilustran el objetivo del muestreo balanceado.

Ejemplo 15.1.1. ***Muestreo aleatorio simple***: esta clase de diseños de muestreo de tamaño fijo n son balanceados sobre la variable $x_k = \pi_k$, $k \in U$. Pues,

$$\sum_{k \in \boldsymbol{S}} \frac{xk}{\pi_k} = \sum_{k \in \boldsymbol{S}} 1 = n = \sum_{k \in U} \pi_k$$

Ejemplo 15.1.2. ***Estratificación***: suponga que en una población estratificada en H estratos (U_h, $h = 1, ..., H$, $\#U_h = N_h$) se selecciona una muestra aleatoria simple de tamaño n_h en cada estrato. El diseño es balanceado sobre las variables

$$\delta_{kh} = \begin{cases} 1 & \text{si la unidad } k \text{ está en el estrato h,} \\ 0 & \text{en otro caso} \end{cases}$$

Puesto que,

$$\sum_{k \in \boldsymbol{S}} \frac{\delta_{kh}}{\pi_k} = \sum_{k \in \boldsymbol{S}} \delta_{kx} \frac{N_h}{n_h} = N_h = \sum_{k \in U} \delta_{kh}$$

En la mayoría de problemas prácticos, las ecuaciones de balance no pueden ser exactamente satisfechas, en otras palabras existe un problema de redondeo que se da porque el inverso de la probabilidad de inclusión no es un entero. Por esta razón, el objetivo es construir un diseño muestral que satisfaga las ecuaciones de balanceo exactamente, si es posible, ó encontrar la mejor aproximación, si no lo es. El problema de redondeo es despreciable cuando el tamaño de muestra esperado es grande.

15.2 El método del cubo

Este método se compone de dos fases, llamadas la fase de vuelo y fase de aterrizaje. En la primera, para que las restricciones sean satisfechas exactamente, se deben redondear a cero (0) o uno (1) las probabilidades de inclusión. La fase de aterrizaje consiste en el manejo adecuado del redondeo.

Como hemos visto, cada vector $\mathbf{s}$, en muestreo sin reemplazo, es un vértice de un N-cubo y el número de posibles muestras es el número de vértices del N-cubo. Un diseño muestral con vector de probabilidades de inclusión $\boldsymbol{\pi}$, consiste en la asignación de una probabilidad a cada vértice.

Geométricamente, un diseño muestral consiste en expresar el vector $\boldsymbol{\pi}$ como una combinación lineal convexa de los vértices del N-cubo. Un algoritmo puede ser visto como un camino (aleatorio) que lleve a alcanzar un vértice del N-cubo de tal manera que se satisfagan las ecuaciones de balanceo.

15.2.1 Fase de vuelo

Es una caminata aleatoria que comienza con un vector de probabilidades de inclusión y permanece en la intersección del cubo y el subespacio restringido por las ecuaciones de balanceo. Esta caminata aleatoria se detiene en un vértice de dicha intersección.

El objetivo de esta fase es escoger aleatoriamente un vértice de

$$K = \{[0,1]^N \cap Q\},$$

donde $Q = \boldsymbol{\pi} + \ker \mathbf{A}$ y $\mathbf{A} = (\check{\mathbf{x}}_1, ..., \check{\mathbf{x}}_N)$, de tal forma que las ecuaciones de balance se reproduzcan a satisfacción. La fase de aterrizaje es sólo necesaria si el vector escogido no es un vértice del cubo y consiste en flexibilizar las restricciones (lo menos posible) para seleccionar una muestra, esto es, un vértice del cubo.

Ejemplo 15.2.1. La fase de vuelo transforma un vector de probabilidades de inclusión en un vector de ceros y unos.

$$\boldsymbol{\pi} = \begin{pmatrix} 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \end{pmatrix} \rightarrow \begin{pmatrix} 0.666 \\ 0.666 \\ 0.666 \\ 0 \end{pmatrix} \rightarrow \begin{pmatrix} 1 \\ 0.5 \\ 0.5 \\ 0 \end{pmatrix} \rightarrow \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix}$$

Si existe un problema de redondeo, entonces algunos componentes no pueden ser convertidos en cero

$$\boldsymbol{\pi} = \begin{pmatrix} 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \end{pmatrix} \rightarrow \begin{pmatrix} 0.625 \\ 0 \\ 0.625 \\ 0.625 \\ 0.625 \end{pmatrix} \rightarrow \begin{pmatrix} 0.5 \\ 0 \\ 0.5 \\ 1 \\ 0.5 \end{pmatrix} \rightarrow \begin{pmatrix} 1 \\ 0 \\ 0.5 \\ 1 \\ 0 \end{pmatrix}$$

15.2.2 La martingala balanceada

El algoritmo general para llevar a cabo la fase de vuelo se realiza utilizando la siguiente definición.

Definición 15.2.1. *Un proceso aleatorio discreto* $\boldsymbol{\pi}(t) = [\boldsymbol{\pi}_k(t)]$ *en* $\mathbb{R}^N$, $t = 0, 1, ...$ *se llama una* ***martingala balanceada*** *para un vector de probabilidades de inclusión* $\boldsymbol{\pi}$ *y para las variables auxiliares* $x_1, ..., x_p$, *si*

1. $\boldsymbol{\pi}(0) = \boldsymbol{\pi}$,
2. $E[\boldsymbol{\pi}(t)|\boldsymbol{\pi}(t-1), ..., \boldsymbol{\pi}(0)] = \boldsymbol{\pi}(t-1)$, $t = 1, 2, ...$
3. $\boldsymbol{\pi}(t) \in K = \{[0,1]^N \cap (\boldsymbol{\pi} + \ker A)$

15.2.3 Implementación de la fase de vuelo

Primero, inicializamos por $\boldsymbol{\pi}(0) = \boldsymbol{\pi}$. Luego, En la etapa $t = 1, ..., T$,

1. Definimos un vector $\mathbf{u}(t) = [u_k(t)] \neq 0$ tal que
   - $\mathbf{u}(t)$ es en el kernel de la matriz A,
   - $u_k(t) = 0$ si $\pi_k(t)$ es entero.
2. Calculamos $\lambda_1^*(t)$ y $\lambda_2^*(t)$, el valor más grande tal que

$$0 \leq \boldsymbol{\pi}(t) + \lambda_1^*(t)u(t) \leq 1,$$

$$0 \leq \boldsymbol{\pi}(t) - \lambda_2^*(t)u(t) \leq 1,$$

3. Elegimos

$$\boldsymbol{\pi}(t) = \begin{cases} \boldsymbol{\pi}(t-1) + \lambda_1^*(t)\mathbf{u}(t) & \text{con probabilidad } q_1(t) \\ \boldsymbol{\pi}(t-1) - \lambda_2^*(t)\mathbf{u}(t) & \text{con probabilidad } q_2(t) \end{cases}$$

donde

$$q_1(t) = \lambda_2^*(t)/(\lambda_1^*(t) + \lambda_2^*(t))$$

y

$$q_2(t) = \lambda_1^*(t)/(\lambda_1^*(t) + \lambda_2^*(t))$$

15.2.4 La fase de aterrizaje

Al final de la primera fase, la martingala balanceada ha alcanzado un vértice de K, el cual no es necesariamente un vértice de C. Este vértice es denotado como $\boldsymbol{\pi}^* = [\pi_k^*] = \boldsymbol{\pi}(T)$. Sea q el número de componentes no enteras en este vértice. Si $q = 0$, el algoritmo está completo. Si $q > 0$ algunas restricciones no pueden ser satisfechas rigurosamente.

Sea $U = \{k \in U|0 < \pi_k^* < 1\}$. El objetivo es buscar un diseño muestral que arroje una muestra $s^* \subset U^*$ tal que

$$\sum_{k\in S} a_k \approx \sum_{k\in U} a_k\pi_k^* = \sum_{k\in U} a_k\pi_k,$$

con $a_k = \check{\mathbf{x}}_k$ y $s^* = s \cap U^*$.

Esto se resuelve mediante programación lineal. Aplicando el método *simplex* tenemos

$$\min_{p^*(\cdot)} \sum_{s^* \subset U^*} Costo(s)p^*(s),$$

sujeto a

$$\sum_{s^* \subset U} p(s^*) = 1 \tag{15.2.1}$$

$$\sum_{s^* \ni k} p(s^*) = \pi_k \tag{15.2.2}$$

$$0 \leq p(s^*) \leq 1 \tag{15.2.3}$$

En donde $Costo(s)$ es el costo de la muestra, que aumenta si las ecuaciones de balanceo, dadas en las secciones anteriores, no se tienen. Luego se selecciona una muestra con un diseño de muestreo $p(\cdot)^*$. Este programa no depende del tamaño poblacional sino sólo del número de variables de balanceo. Si el número de variables auxiliares es muy grande, al final de la fase de vuelo se debe eliminar una variable auxiliar. Por esta razón es importante ordenar las variables de balanceo de acuerdo a la correlación con las variables de interés.

Varianza

Deville & Tillé (2005) han propuesto aproximar la varianza suponiendo que el muestreo balanceado se puede suponer como un muestreo condicional de Poisson. Así,

$$Var(\hat{t}_{y,\pi}) = Var(\hat{E}_{poisson}) = \frac{N}{N-p} \sum_{k \in U} \frac{E_k^2}{\pi_k^2} \pi_k(1-\pi_k), \tag{15.2.4}$$

donde $E_k = y_k - \mathbf{x}_k'\mathbf{B}$.

Ejemplo 15.2.2. Nótese que la misma función que cumple el muestreo balanceado, la cumple el diseño de muestreo πPT, puesto que, en virtud del conocimiento de un característica de interés, se garantiza, siguiendo el resultado 4.3.2, que el estimador del total poblacional de la característica de información auxiliar, $\hat{t}_{x,\pi}$, reproduzca al total poblacional de la característica de interés, t_x, con varianza nula.

Sin embargo, el diseño de muestreo πPT, cumple esta función solamente para una y sólo una característica de información auxiliar, y cuando el investigador puede tener acceso a varias características de información auxiliar de manera simultanea, entonces el muestreo πPT deja de ser útil. En este orden de ideas, se pude decir que, abusando del lenguaje, el diseño de muestreo balanceado es una generalización del diseño de muestreo πPT.

Este ejemplo trata de ilustrar el procedimiento computacional para la obtener el objetivo final de la selección de una muestra balanceada. Se utilizará la población `MU284` (Särndal, Swensson & Wretman 1992) para tales efectos. En primer lugar suponga, sin pérdida de generalidad, que se planea utilizar, en principio, un diseño de muestreo πPT (podría ser cualquier otro diseño de muestreo). Utilizando

la función `inclusionprobabilities` del paquete `sampling`, se obtienen las probabilidades de inclusión inducidas por este diseño de muestreo con probabilidad proporcional a la característica de información auxiliar P75. Nótese que el tamaño de la muestra es de 50 unidades.

```
> library(sampling)
> data(MU284)
> attach(MU284)
> pik=inclusionprobabilities(MU284$P75,50)
> sum(pik)
[1] 50
```

Suponga que deseamos obtener una muestra balanceada con respecto a todas las características de información auxiliar dadas por P75, CS82, SS82, S82, ME84 Y REV84. Para esto, incluimos todos los valores poblacionales observados de estas variables de balanceo en una matriz. A continuación, utilizamos la función `samplecube` para obtener una muestra que sea balanceada con respecto a todos los totales poblacionales de todas las variables de balanceo.

```
> X=cbind(MU284$P75,MU284$CS82,MU284$SS82,
          MU284$S82,MU284$ME84,MU284$REV84)
> s=samplecube(X,pik,order=1,comment=TRUE)

BEGINNING OF THE FLIGHT PHASE
The matrix of balancing has 6 variables and  284  units
The size of the inclusion probability vector is  284
The sum of the inclusion probability vector is  50
The inclusion probability vector has 281 non-integer elements
Step 1

BEGINNING OF THE LANDING PHASE
At the end of the flight phase, there remain 6 non integer
probabilities. The sum of these probabilities is  2
This sum is  integer
The linear program will consider  15  possible samples
The mean cost is  0.02644766
The smallest cost is  0.005041543
The largest cost is  0.05423445
The cost of the selected sample is 0.01280451

QUALITY OF BALANCING
  TOTALS HorvitzThompson_estimators Relative_deviation
1   8182                   8182.000      -5.557900e-14
2   2583                   2567.547      -5.982456e-01
3   6301                   6194.726      -1.686628e+00
4  13500                  13275.528      -1.662755e+00
5 505226                 503726.047      -2.968875e-01
6 874017                 881901.409       9.020887e-01
```

Nótese que la salida de esta función es muy explicativa. Para este caso particular,

se necesito tanto de la fase de vuelo como de la fase de aterrizaje. Al final de la fase de vuelo, quedaban seis individuos cuyas probabilidades no eran cero o uno. Por lo tanto, el método del cubo, necesita de la fase de aterrizaje para alcanzar convergencia. Además de los comentarios para cada fase del método del cubo, esta función también devuelve una tabla que describe la calidad del procedimiento en términos de la desviación relativa. El lector no debe pasar por alto la calidad del balanceo. Es simplemente extraordinario que se consiga tal exactitud con una muestra de tan sólo 50 unidades.

15.3 Marco y Lucy

Este capítulo cierra con la implementación del método del cubo para la selección de muestras balanceadas. Suponga que el investigador conoce el comportamiento estructural de algunas características de interés; a saber, Ingreso y Número de empleados. Para seleccionar una muestra balanceada, en principio, fijas las probabilidades de inclusión de acuerdo a un diseño de muestreo aleatorio simple. Como de costumbre, inserta la matriz de observaciones de las características de interés en la función `samplecube`.

```
> library(TeachingSampling)
> data(Lucy)
> attach(Lucy)
> pik=rep(400/2396,2396)
> X <- cbind(Income, Employees)
> s=samplecube(X,pik,order=1,comment=TRUE)

BEGINNING OF THE FLIGHT PHASE
The matrix of balanced variable has 2  variables and  2396  units
The size of the inclusion probability vector is  2396
The sum of the inclusion probability vector is  400
The inclusion probability vector has  2396  non-integer elements
Step 1

BEGINNING OF THE LANDING PHASE
At the end of the flight phase, there remain
2 non integer probabilities
The sum of these probabilities is  1.046484
This sum is  non-integer
The linear program will consider  3  possible samples
The mean cost is  0.001172141
The smallest cost is  6.929445e-05
The largest cost is  0.002644001
The cost of the selected sample is 6.929445e-05

QUALITY OF BALANCING
           TOTALS   HorvitzThompson  Relative_deviation
Income    1035217         1035754.9         0.051956257
Employees  151950          151954.3         0.002843040
```

Para este caso particular, la función `samplecube` que implementa el método del cubo necesitó tanto de la fase de vuelo como de la fase de aterrizaje para alcanzar la convergencia. La fase de vuelo concluyó con 2 elementos cuyas probabilidades de inclusión no eran cero o uno. Sin embargo, después de la fase de aterrizaje una muestra balanceada fue seleccionada. Una vez más no puede pasar inadvertida la calidad del balanceo.

Después de haber seleccionado la muestra balanceada, es tiempo de obtener las estimaciones pertinentes. En general, es posible utilizar la función `E.piPS` del paquete `TeachingSampling` puesto que el marco general del muestreo balanceado se acomoda a las características que rigen la estimación de Horvitz-Thompson.

```
> sam <- (1:length(pik))[s==1]
> pik.s <- pik[sam]
> muestra <- Lucy[sam,]
> attach(muestra)
> estima <- data.frame(Income, Employees, Taxes)

> E.piPS(estima,pik.s)
                 Income    Employees        Taxes
Estimation 1.035755e+06 1.519543e+05 2.957862e+04
Variance   8.739838e+08 1.253370e+07 3.521523e+06
CVE        2.854268e+00 2.329843e+00 6.344353e+00
```

Los resultados que arroja la función son óptimos, en el sentido de que además de obtener estimaciones cercanas al total poblacional para la característica de interés también mantiene los totales poblaciones de las características de interés en el diseño de muestreo.

Tabla 15.1: *Muestreo balanceado: estimación de los totales de las características de interés.*

Variable	Total poblacional	Total estimado	cve %	Desv. %
Ingreso	1035217	1035755	2.85	0.05
Empleados	151950	151954	2.32	0.00
Impuestos	28654	29578	6.34	3.22

15.4 Desarrollos recientes y preguntas frecuentes

El muestreo balanceado es un procedimiento común en el presente. Se utiliza para la selección de la muestra maestra en el INSEE, en Francia y para el nuevo censo Frances que selecciona cada año un quinto de la población para entrevistarla. Es decir en cinco la población es censada.

15.4.1 Algunas preguntas

Tillé (2006) responde algunas preguntas que surgen directamente con respecto al funcionamiento de este nuevo método en la práctica:

- **¿Por qué no usar calibración en vez de balanceo?**
La estratificación es un caso particular del muestreo balanceado, la post-estratificación es un caso particular de la calibración. En estratificación y balanceo, los pesos no son aleatorios. Esto hace que sea una mejor estrategia. La calibración tiene la ventaja de sólo requerir el conocimiento de los totales poblacionales de las variables auxiliares, mientras que en el balanceo se require el conocimiento de los valores de las variables auxiliares para todas las unidades de la población.

- **¿Qué tan precisa es la aproximación de la estimación en muestreo balanceado?**
Deville & Tillé (2004) han comprobado que bajo condiciones de regularidad realistas en la vida práctica se tiene que

$$\left|\frac{\hat{t}_{x_q,\pi} - t_{xq}}{t_{xq}}\right| < O(p/N) \leq o_p(\sqrt{1/N})$$

para todo $q = 1, \ldots, Q$.

- **¿Cómo estimar la varianza?**
Mediante una técnica de residual desarrollada en Deville & Tillé (2005). Esta técnica es comparable con la técnica usada para calcular la varianza del estimador de calibración y ha sido validada mediante un conjunto de simulaciones.

- **¿Se puede usar balanceo y calibración simultáneamente?**
Ambas técnicas pueden ser usadas juntas. No hay ninguna contradicción. La mejor estrategia muestral consistiría en usarlas juntas. De hecho la calibración puede arreglar el problema del redondeo después del balanceo. Más aún, se pueden utilizar distintas variables en la calibración de las usadas en el balanceo.

- **¿Qué software usar?**
En SAS-IML, existen dos paquetes (INSEE y University of Neuchâtel), en R el paquete `sampling` permite usar el método del cubo. Estos softwares están disponibles en internet de manera gratuita.

15.5 Ejercicios

15.1 Suponga un diseño de muestreo de tamaño $n = 2$ para una población de tamaño $N = 3$ con una característica de información auxiliar tal que $x_k = \pi_k$ (k=1,2,3) y además $\pi_1 + \pi_2 + \pi_3 = 2$

- Escriba las ecuaciones de balanceo.
- Calcule las entradas de la matriz $\mathbf{A}$ (sección 15.2.1).
- Defina el espacio nulo de la matriz $\mathbf{A}$; es decir $\ker(\mathbf{A})$.
- Obtenga la forma explícita de $Q = \boldsymbol{\pi} + \ker(\mathbf{A})$.

15.2 Suponga un diseño de muestreo balanceado con $N = 8$ y $n = 4$. Asuma que, el vector de probabilidades de inclusión de primer orden es

$$\boldsymbol{\pi} = \left(\frac{1}{9}, \frac{2}{9}, \frac{3}{9}, \frac{4}{9}, \frac{5}{9}, \frac{6}{9}, \frac{7}{9}, \frac{8}{9}\right)'$$

y existen dos variables de balanceo; la primera, $x_{1k} = \pi_k$ y la segunda, $x_{2k} = 1$, para todo $k \in U$.

- Escriba las ecuaciones de balanceo.
- Calcule las entradas de la matriz $\mathbf{A}$.
- Si la función de costo es

$$Costo_1(\mathbf{s}) = \sum_{p=1}^{P} \frac{(\hat{t}_{x_p,\pi} - t_x)^2}{t_x^2}$$

  Obtenga el costo generado por la fase de aterrizaje para las muestras:
  $\mathbf{s}_1 = (1, 0, 0, 0, 0, 1, 1, 1)'$.
  $\mathbf{s}_2 = (0, 0, 0, 1, 1, 1, 0, 1)'$.
  $\mathbf{s}_3 = (0, 0, 1, 1, 0, 0, 1, 1)'$.
  $\mathbf{s}_4 = (0, 0, 1, 1, 0, 1, 1, 0)'$.

- Si la función de costo es

$$Costo_2(\mathbf{s}) = (\mathbf{s} - \boldsymbol{\pi})'\mathbf{A}'(\mathbf{A}\mathbf{A}')^{-1}\mathbf{A}(\mathbf{s} - \boldsymbol{\pi})$$

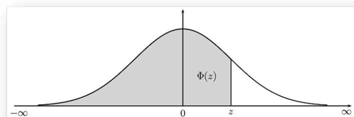
  Obtenga el costo generado por la fase de aterrizaje para las anteriores muestras.

15.3 Demuestre o refute las siguientes afirmaciones

- «Utilizar muestreo balanceado siempre mejora la eficiencia de la estrategia de muestreo»
- «Utilizar calibración siempre mejora la eficiencia de la estrategia de muestreo balanceado»
- «Utilizar calibración y muestreo balanceado siempre mejora la eficiencia de la estrategia de muestreo»

Apéndice A

Distribución normal estándar



Cuantil z	Probabilidad acumulada $\Phi(z)$	Cuantil z	Probabilidad acumulada $\Phi(z)$
-1.959	0.025	0.062	0.525
-1.644	0.050	0.125	0.550
-1.439	0.075	0.189	0.575
-1.281	0.100	0.253	0.600
-1.150	0.125	0.318	0.625
-1.036	0.150	0.385	0.650
-0.934	0.175	0.453	0.675
-0.841	0.200	0.524	0.700
-0.755	0.225	0.597	0.725
-0.674	0.250	0.674	0.750
-0.597	0.275	0.755	0.775
-0.524	0.300	0.841	0.800
-0.453	0.325	0.934	0.825
-0.385	0.350	1.036	0.850
-0.318	0.375	1.150	0.875
-0.253	0.400	1.281	0.900
-0.189	0.425	1.439	0.925
-0.125	0.450	1.644	0.950
-0.062	0.475	1.959	0.975
0.000	0.500	2.033	0.979

Bibliografía

Apostol, T. M. (1963), *Mathematical Analysis*, Adison Wesley.

Ardilly, P. (1991), 'Échantillonnage représentatif optimum à probabilités inégales', *Annales d'Économie et de Statistique* **23**, 91–113.

Basu, D. (1971), 'An essay on the logical foundations of survey sampling', *Holt, Rinehart and Winston. Toronto* pp. 203–242.

Bautista, J. (1998), *Diseños de muestreo estadístico*, Universidad Nacional de Colombia.

Bebbington, A. (1975), 'A simple method of drawing a sample without replacement', *Applied Statistics* **24**, 136.

Bethelehem, J. & Keller, W. A. (1987), 'Linear weighting of survey data.', *Journal of Official Statistics* **3**, 141–153.

Binder, D. (1983), 'On the variances of asymptotically normal estimators from complex surveys', *International Statistical Review* **51**, 279–292.

Bowley, A. L. (1906), 'Address to the economic science and statistics section of the british association for the advancement of sciences', *Journal of the Royal Statistical Society* **69**, 548–557.

Bowley, A. L. (1926), 'Measurement og the precision attained in sampling', *Bulletin of the International Statistical Institute* .

Box, G. E. P. (1979), *Robustness in the Strategy of Scientific Model Building in Robusteness in Statistics*, Academis Press.

Breidt, F. & Opsomer, J. D. (2000), 'Local polynomial regression estimators in survey sampling', *The Annals of Statistics* **28**, 1026–1053.

Brewer, K. (1963), 'A model of systematic sampling with unequal probabilities', *Australina Journal of Statistics* **5**, 93–105.

Brewer, K. (1975), 'A simple procedure for πpswor', *Australian Journal of Statistics* **17**, 166–172.

Brewer, K. (2002), *Combined sampling inference, weighting Basu's elephants*, London: Arnorld.

Brewer, K. & Hanif, M. (1983), *Sampling with unequal probabilities*, New York: Springer-Verlag.

Brewer, K. R. W. (1999), 'Cosmetic calibration for unequal probability samples', *Survey Methodology* **25**, 205–212.

Cassady, R. & Valliant, R. (1993), 'Condicional properties of poststratified estimators under normal theory', *Survey Methodology* **19**, 183–192.

Cassel, C., Särndal, C. & Wretman, J. (1976*a*), *Foundations of Inference in Survey Sampling*, Wiley.

Cassel, C., Särndal, C. & Wretman, J. (1976*b*), 'Some results on generalized difference estimation and generalized regression estimation for finite populations', *Biometrika* **63**, 615–620.

Chambers, R. L. & Skinner, C. J., eds (2003), *Analysis of Survey Data*, Wiley.

Chen, J. & Qin, J. (1993), 'Empirical likelihood estimation for finite populations and the efectivene usage of auxiliary information', *Biometrika* **80**, 107–116.

Cochran, W. (1977), *Sampling Techniques*, Wiley.

Cochran, W. G. (1939), 'The use of analysis of variance in enumeration by sampling', *Journal of the American Statistical Association* **34**, 492–510.

Cochran, W. G. (1940), 'The estimation of the yields of cereal experiments by sampling for the ratio of grain to total produce', *Journal of Agricultural Science* **30**, 262–275.

Cornfield, J. (1951), 'The determination of sampling size', *American journal of public health* **41**, 654–661.

Dalgaard, P. (2008), *Introductory Statistics with R*, 2 edn, Springer.

Deming, W. & Stephan, F. (1940), 'On a least squares adjustment of a sampled frequency table when the expected marginal totals are known', *Annals of Mathematical Statistics* **11**, 427–444.

Deville, J. (1993), 'Estimation de la variance pour les enquetes en deux phases', *Note Interne Manuscrite. France: INSEE* .

Deville, J. C. (1992), Constrained samples , conditional inference, weihting: Three aspects of the utilisation of auxiliary information, *in* S. Örebro, ed., 'Proceedings of the Workshop on the Uses of Auxiliary Information in Survey'.

Deville, J.-C. (1999), 'Variance estimation for complex statistics and estimators: linearizaion and residual techniques', *Survey Methodology* **25**, 193–204.

Deville, J.-C., Särndal, C.-E. & Sautory, O. (1993), 'General raking procedures in survey sampling', *Journal of the American Statistical Association* **88**, 1013–1020.

Deville, J.-C. & Tillé, Y. (1998), 'Unequal probability sampling without replacement through a splitting method', *Biometrika* **85**, 89–101.

Deville, J.-C. & Tillé, Y. (2005), 'Variance approximation under balanced sampling', *Journal of Statistical Planning and Inference* **128**, 411–425.

Deville, J. C. & Tillé, Y. (2004), 'Efficient balanced sampling: The cube method', *Biometrika* **91**, 893–912.

Deville, J. & Särndal, C. (1992), 'Calibration estimators in survey sampling', *Journal of the American Statistical Association* **87**, 376–382.

Draper, D. (1998), 'Rank-based robust analysis of linear models i. exposition and review', *Statistical Science* **3**, 239–257.

Durbin, J. (1967), 'Design of multi-stage surveys for the estimation of sampling errors', *Applied statistics* **16**, 152–164.

Estevao, V. M. & Särndal, C.-E. (2001), 'The ten cases of auxiliary information for calibration estimators in two-phase sampling', *Journal of Official Statistics* **18**, 233–255.

Estevao, V. M., Särndal, C.-E. & Sautory, O. (2000), 'A functional form approach to calibration', *Journal of Official Statistics* **16**, 379–399.

Fan, C., Muller, M. & Rezucha, I. (1962), 'Development of sampling plans by using sequential (item by item) selection techniques and digital computer', *Journal of the American Statistical Association* **57**, 387–402.

Frankel, M. & King, B. (1996), 'A conversation with leslie kish', *Statistical Science* **11**, 65–87.

Gelman, A., Carlin, B. C., Stern, H. S. & Rubin, D. (2004), *Bayesian Data Analysis*, 2 edn, Chapman & Hall / CRC.

Godambe, V. P. (1955), 'A unified theory of sampling for the finite populations', *Journal of the Royal Statistical Society. Series B* **17**, 73–96.

Godambe, V. P. & Thompson, M. E. (1977), 'Robust near optimal estimation in survey practice', *Bulletin of the International Statistical Institute* .

Groves, R., Fowler, F., Couper, M., Lepkowski, J., Singer, E. & R., T. (2004), *Survey Methodology*, Wiley.

Gupta, A. K. & Nagar, D. K. (1999), *Matriz Variate Distributions*, Chapman and Hall.

Gutierrez, H. A. (2009), *TeachingSampling: Sampling designs and parameter estimation in finite population.* R package version 0.7.6.

Gutiérrez, H. A. (2009*a*), 'Estimation stage in survey sampling: A multiparameter approach', *Revista Colombiana de Estadística* **32**, 79–97.

Gutiérrez, H. A. (2009*b*), 'Model assisted survey regression estimators: A rank-based approach', *Proceedings of the 57th Session of the International Statistical Institute* pp. 18–31.

Gutiérrez, H. A. & Breidt, F. J. (2009), 'Estimation of the population total using the generalized difference estimator and wilcoxon ranks', *Revista Colombiana de Estadística* **32**, 123–143.

Hájek, J. (1960), 'Limiting distributions in simple random sampling from a finite poulation', *Publication of Mathematical Institute of the Hungarian Academy of Science* **5**, 361–374.

Hájek, J. (1971), 'Comment on an essay on the logical foundations of survey sampling, part one', *The Foundations of Survey Sampling* pp. Godambe, V.P. and Sprott, D.A. eds., 236, Holt, Rinehart, and Winston.

Hájek, J. (1981), *Sampling from a finite population*, New York: Marcel Dekker.

Hansen, H. M. & Hurwitz, W. N. (1943), 'On the theory of sampling from finite populations', *Annals of Mathemathical Statistics* **14**, 333–362.

Hansen, M., Hurwitz, W. & Madow, W, G. (1953), *Sample survey methods and theory. Vols. I and II*, John Wiley and Sons.

Hartley (1959), 'Analytic studies of survey data', *Instituto di Statistica* **Volume in honor of Corrado Gini**.

Hettmansperger, T. (1984), *Statistical Inference Based on Ranks*, Wiley. New York.

Hettmansperger, T. & McKean, J. (1998), *Robust Nonparametric Statistical Methods*, Great Britain: Arnold.

Hidiroglou, M. A. & Rao, J. N. K. (2003), Variance estimation in two-phase sampling, *in* S. Canada, ed., 'Proceedings of Statistics Canada Symposium', pp. 2–13.

Holmberg, A. (2002*a*), 'A multiparameter perspective on the choice of sampling design in surveys', *Statistics in Transition* **5**, 969–994.

Holmberg, A. (2002*b*), On the choice of sampling design under greg estimation in multiparameter surveys, Technical Report 1, *RD Department, Statistics Sweden*, SE-701 89 Örebro, Sweden.

Holt, D. & Smith, T. (1979), 'Post-stratification.', *Journal of the Royal Statistical Society* **A142**, 33–46.

Horvitz, D. & Thompson, D. (1952), 'A generalization of sampling without replacement from a finite universe', *Journal of the American Statistical Association* **47**, 663–685.

Hubback, J. A. (1927), 'Sampling for rice yield in bihar and orissa', *Imperial Aggricultural Research Institute* **166**.

Isaki, C. & Fuller, W. A. (1982*a*), 'Survey design under the regression superpopulation model', *Journal of the American Statistical Association* **767**, 89–96.

Isaki, C. T. & Fuller, W. A. (1982*b*), 'Survey design under the regression superpoluation model', *Journal of the American Statistical Association* **77**, 89–96.

Jaeckel, L. (1972), 'Estimating regression coefficients by minimizing the dispersion of the residuals', *The Annals of Mathematical Statistics* **43**, 1449–1458.

Jagers, P. (1986), 'Post-estratification against bias in sampling', *International Statistical Review.* **54**, 159–167.

Kiaer, A. N. (1901), 'Sur les méthodes representatives ou typologiques', *Bulletin of the International Statistical Institute* .

Kish, L. (1965), *Survey Sampling*, Wiley.

Kott, P. S., Sweensson, B., Särndal, C. E. & Wretman, J. (2005), 'An interview with the authors of the book: Model-assisted survey sampling', *Journal of Official Statistics* **21**, 171–182.

Kyburg, H. E. (1987), 'the basic bayesian blunder', *The Foundations of Statistical Inference* pp. Reidel, Dordrecht.

Lahiri, D. (1951), 'A method for sample selection providing unbiased ration estimates', *Bulletin of the International Statistical Institute.* **33,2**, 133–140.

Lahiri, D. B. (1968), 'On the unique sample, the surveyed one', *Foundations of Survey Sampling* .

Laplace, P. (1814), *A Philosophical Essay on Probability.*

Lavallée, P. (2007), *Indirect Smapling*, Springer.

Lavallée, P. & Caron, P. (2008), 'Estimation using the generalized weigth share method: The use of record linkage', *Survey Methodology* **27**, 155–169.

Lehtonen, R. & Pahkinen, E. (2003), *Practial methods for design and analysis of complex surveys*, 2 edn, New York: Wiley.

Lohr, S. (2000), *Sampling: Design and Analysis*, Thompson.

Madow, W. (1948), 'On the limiting distributions based on samples from finite universes', *Annals of Mathematical Statistics* **19**, 535–545.

Madow, W. G. & Madow, L. L. (1944), 'On the theory of systematic sampling', *Annals of Mathematical Statistics* **15**, 1–24.

Mahalanobis, P. (1946), 'Recent experiment in statistical sampling in the indian statitical institute', *Journal of the Royal Statistical Society* **109**, 325–370.

Matei, A. & Tille, Y. (2005), 'Evaluation of variance approximations and estimators in maximun entropy sampling with unequal probability and fixed sample size', *Journal of Official Statistics.* **4**, 543–570.

McCullagh, P. & Nelder, J. A. (1989), *Generalized Linear Models*, Chapman and Hall.

Mood, A. M., Graybill, F. A. & Boes, D. C. (1974), *Introduction to the Theory of Statistics*, 3 edn, McGraw Hill.

Narain, R. (1951), 'On sampling without replacement with varying probabilities', *Journal of Indian Society of Agricultural Statistics* **3**, 169–175.

Neyman, J. (1934), 'On the two differents aspects of the representative method: the method of stratified sampling and the method of purposive selection', *Journal of the Royal Statistical Society* **97**, 558–625.

Neyman, J. (1938), 'Contribution to the theory of sampling human populations', *Journal of the American Statistical Association* **33**, 101–116.

Ospina, D. (2001), *Introducción al muestreo.*, Universidad Nacional de Colombia.

R Development Core Team (2008), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Raj, D. (1954), 'On sampling with probabilities proportional to size', *Ganita* **5**, 175–182.

Raj, D. (1968), *Sampling theory*, McGraw Hill.

Rao, J. N. K. (1973), 'On double sampling for stratification and analityc surveys', *Biometrika* **60**, 125–133.

Rao, J. N. K. (2005), 'Interplay between survey theory and practice: An appraisal', *Survey Methodology* **31**, 117–138.

Ravishanker, N. & Dey, D. (2002), *A First Course in Linear Model Theory*, Chapman and Hall.

Rosén, B. (1972), 'Asymptotic theory for succesive sampling with varying probabilities without replacement, i and ii', *Annals of Mathematical Statistics* **43**, 373–397, 748–776.

Royal, R. M. & Herson, J. (1973), 'Robust estimation in finite population ii: Estratification on a size variable', *Journal of the American Statistical Association* **68**, 891–893.

Royall, R. M. (1971), 'Linear regression models in finite population samplijng theory', *The Foundations of Survey Sampling* pp. Godambe, V.P. and Sprott, D.A. eds., 236, Holt, Rinehart, and Winston.

Royall, R. M. (1976), 'The linear least squares prediction approach to two stage sampling', *Journal of the American Statistical Association* **71**, 657–664.

Sampath, S. (2001), *Sampling Theory and Methods*, Narosa Publishing House.

Särndal, C. (1980), 'On π-inverse weighting versus best linear unbiased weighting in probability sampling', *Biometrika* **67**, 639–650.

Särndal, C.-E. (2007), 'The calibration approach in survey theory and practice', *Survey Methodology* **33**, 99–119.

Sarndal, C.-E. & Wright, R. L. (1984), 'Cosmetic form of estimators in survey sampling', *Scandinavian Journal of Statistics* **11**, 164–156.

Särndal, C., Swensson, B. & Wretman, J. (1989), 'The weighted residual technique for estimating the variance of the general regression estimator of the finite popoulation total', *Biometrika* **76**, 527–537.

Särndal, C., Swensson, B. & Wretman, J. (1992), *Model Assisted Survey Sampling*, Springer, New York.

Scott, A. J. & Smith, T. M. F. (1969), 'Estimation in multistage surveys', *Journal of the American Statistical Association* **64**, 830–840.

Sen, A. (1953), 'On the estimate of the variance in sampling with varying probabilities', *Journal of the Indian Society of Agricultural Statistics* **5**, 119–127.

Shao, J. (2003), *Mathematical Statistics*, Springer.

Smith, T. (1991), 'Post-estratification', *The Statistician* **40**, 315–323.

Smith, T. M. F. (1976), 'The foundations of survey sampling: a review (with discussion)', *Journal of the Royal Statistical Society* **139**, 183–204.

Srinath, K. P. & Hidiroglou, M. A. (1980), 'Estimation of variance in multi-stage sampling', *Metrika* **27**, 121–125.

Sudgen, R. A. & Smith, T. M. F. (1984), 'Ignorable and informative designs in survey sampling inference', *Biometrika* **71**, 495–506.

Sunter, A. (1977), 'List sequential sampling with equal or unequal probabilities without replacement', *Applied Statistics* **26**, 261–268.

Sunter, A. (1986), 'Solutions to the problem of unequal probabilities sampling without replacement', *International Statitical Review* **54**, 33–50.

Särndal, C. E. & Lundström, S. (2004), *Estimation in Surveys with Nonresponse*, Wiley.

Särndal, C. E. & Sweensson, B. (1987), 'A general view of estimation for two phases of selection with aplications to two-phase sampling and nonresponse', *International Statistical Review* **55**, 279–294.

Terpstra, J. & McKean, J. (2005), 'Rank-based analyses of linear models using `R`', *Journal of Statistical Software* **14**, 1–26.

Tillé, Y. (2006), *Sampling Algorithms*, Springer.

Tille, Y. & Ardilly, P. (2006), *Sampling Methods: Exercises and Solutions*, Springer.

Tillé, Y. & Matei, A. (2008), *sampling: Survey Sampling.* R package version 2.0.

Tippett, L. C. (1927), 'Random sampling numbers', *Tracts for Computers* **XV**.

Valliant, R., Dorfman, A. H. & Royall, R. M. (2000), *Finite Population Sampling and Inference*, Wiley.

Woodruff, R. (1971), 'A simple method for approximating the variance of a complicated estimate', *Journal of the American Statistical Association* **66**, 411–414.

Wu, C. (2003), 'Optimal calibration estimators in survey sampling', *Biometrika* **90**, 937–951.

Wu, C. & Luan, Y. (2003), 'Optimal calibration estimators under two-phase sampling', *Journal of Official Statistics* **19**, 119–131.

Wu, C. & Sitter, R. (2001*a*), 'A model calibration approach to using complete auxiliary information from survey data', *Journal of the American Statistical Association* **96**, 185–193.

Wu, C. & Sitter, R. R. (2001*b*), 'A model-calibration approach to using complete auxiliary information from survey data', *Journal of the American Statistical Association* **96**, 185–193.

Yates, F. (1946), 'A review of recent statistical developments in sampling and sampling surveys', *Journal of the Royal Statistical Society* **A109**, 12–43.

Yates, F. & Grundy, P. (1953), 'Selecting withou replacement from within estrata with probability proportional to size', *Journal of the Royal Statitical Society* **B15**, 235–261.

Zhang, L. (2000), 'Post-estratification and calibration - a synthesis', *The American Statistician.* **54**, 178–184.

Índice de figuras

Índice de Tablas

Índice alfabético

El término Estrategia de Muestreo no ha tenido la trascendencia pertinente en el mundo del muestreo. Se habla de la precision e incluso insesgamiento de un estimador sin tener en cuenta que tales propiedades están ligadas al diseño de muestreo que se haya utilizado en la recoleccion de la información. Para el autor, el aprendizaje de esta materia es más sencillo cuando se valora de igual manera el diseño de muestreo junto con el estimador del parámetro de interés utilizado en la población finita. No se puede desconocer la regla de oro del muestreo que clama: utilizar diseños de muestreo que induzcan probabilidades de inclusión (o selección, según sea el caso) proporcionales al valor de la característica de interés en la población y utilizar estimadores que involucren dichas probabilidades. Por lo anterior, se ha decidido titular este texto como Estrategias de muestreo, diseño de encuestas y estimación de parámetros.

En la búsqueda combinada de una mejor estrategia de muestreo se ha querido dividir este texto en cuatro partes que pueden ser utilizadas en los distintos niveles de pregrado así como en cursos de posgrado, dependiendo de la dicultad del tema. La división del libro corresponde al desarrollo teórico del muestreo a través de su corta historia: *recuento de las estrategias de muestreo más utilizadas; utilización y aprovechamiento de la información auxiliar disponible en el marco del muestreo; avances metodológicos en el muestreo; y, estimadores de calibración, el muestreo balanceado y el muestreo indirecto.*

- **Use la estrategia pedagógica de Marco y Lucy**
- **Uso del paquete TeachingSampling**
- **Ejemplos prácticos**
- **Ejercicios propuestos**

H. Andrés **Gutiérrez** Rojas



Es uno de los estadísticos más influyentes en Colombia. Autor de varios libros en estadística y de diversos artículos en revistas indexadas internacionales. Conferencista invitado en eventos nacionales e internacionales. Experto en muestreo, diseño y análisis de encuestas, estadística bayesiana y metodologías de evaluación. Entre otros, fue profesor de la Universidad Nacional de Colombia, asesor en muestreo para la Dirección de Seguimiento y Evaluación de Políticas Públicas del DNP (2010 - 2015), decano de la Facultad de Estadística de la USTA (2013 -2015). Desde 2015 se desempeña como Director Técnico de Evaluación en el ICFES y fue recientemente nombrado consejero científico del OCyT.



Contenidos libres en:


www.edicionesdelau.com



ISBN 978-958-762-586-8